AD_______________

Award Number: DAMD17-03-1-0034

TITLE: Short- and Long-Term Effects in Prostate Cancer Survival: Analysis of Treatment Efficacy and Risk Prediction

PRINCIPAL INVESTIGATOR: Alexander D. Tsodikov, Ph.D.

CONTRACTING ORGANIZATION: University of California
Davis, CA 95616-8671

REPORT DATE: March 2006

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
01-03-2006

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*
1 MAR 2003 - 28 FEB 2006

4. TITLE AND SUBTITLE
Short- and Long-Term Effects in Prostate Cancer Survival: Analysis of Treatment Efficacy and Risk Prediction

5a. CONTRACT NUMBER

5b. GRANT NUMBER
DAMD17-03-1-0034

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Alexander D. Tsodikov, Ph.D.

E-mail: atsodikov@ucdavis.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of California
Davis, CA 95616-8671

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
Original contains color plates: All DTIC reproductions will be in black and white.

14. ABSTRACT

All tasks and specific aims of the project have been addressed. A set of methodological tools and software has been developed and applied to analyze prostate cancer data.

15. SUBJECT TERMS
No subject terms provided.

16. SECURITY CLASSIFICATION OF:

a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
U	U	U	UU	120	**19b. TELEPHONE NUMBER** *(include area code)*

Contents

1 Statement of Work

Short- and Long-term Effects in Prostate Cancer Survival: Analysis of Treatment Efficacy and Risk Prediction
Alexander Tsodikov, Ph.D.

In the course of the project there has been no change in the scope of work. All tasks have been completed. A breakdown below shows what has been accomplished.

Task 1. Develop model-building techniques

Task 2. Develop estimation and hypothesis testing

Task 3. Develop variable selection procedures

Task 4. Analysis of the data for significant effects

Task 5. Computer-intensive approaches to prognosis and validation

2 Objectives

There has been no change in the project objectives. The specific aims of this project are

1. To provide a statistical model that reproduces the complex survival responses in prostate cancer.

2. To develop methodology for analysis of prognosis after treatment for prostate cancer taking into account the long- and short-term effects of prognostic factors and treatment.

3. To develop statistical software implementing model-building, estimation, construction of prognostic indices, conditional survival prognosis, and assessment of the quality of prognostic classifications based on the new models.

4. To apply the models and methodology to analyze post-treatment survival of patients with prostate cancer using data from the Memorial Sloan Kettering Cancer Center and the SEER database.

3 Introduction

This project represents a successful effort to develop abstract statistical theory, computational algorithms, translate this methodology into stable shareware software products that can be used by the broad scientific community, put this product into the R-Projects `nltm` and `rpNLTM` that has become the dominant site for dissemination of cutting edge statistical procedures, and finally use and showcase all this arsenal to address real data and problems in prostate cancer. We are glad that we took the challenge of this large idea development and translational effort in the three year project performance period and that were able to see it through.

The goal of this proposal was to investigate a novel approach to the analysis of post-treatment survival of prostate cancer patients: the decomposition of the diversity of survival patterns into short-term and long-term effects. We proposed to identify a model of prostate cancer survival incorporating long- and short-term effects of prognostic factors and treatment. Novel statistical tools were developed to make such models work for better prognosis of prostate cancer patients. Year 1 at the University of Utah was primarily devoted to development of methodology for point estimation and hypothesis testing. While continuing methodological research in Year 2, we focused on the delivery aspect of the progect addressing software development and implementation of the algorithms, testing them by simulations, development of tools for multivariate analysis and variable selection and preliminary applications of these tools to real data. In the last 3rd year of the grant (a no-cost extention), we focused on justifications for asymptotic theory, development of computer-intensive data mining tools using our models, developing an R-package that would give a broad scientific community access to the free software tools implementing the methodology developed in this project, and on applications of these methods to analyze data on survival of patients treated for prostate cancer from Sloan-Kettering Cancer Center Database and SEER database.

In the following sections of the report we give a summary of the results achieved in this project.

4 Methodology

4.1 Models and inference procedures

Motivated by second-order properties of frailty models we have proposed a family of so-called Nonlinear Transformation Models (NTM) (Year 2 report, Section 4). The models were supplied with a general numerical inference framework based on the QEM algorithm (Year 1 report, Section 4). We developed composition techniques that allowed us to easily extend NTMs in a hierarchical fashion into more complex models (Year 1 report, Section 5, Year 2 report Section 7). We proved that the QEM estimation algorithm will fit any model constructed using the techniques. This framework was used to come up with a flexible family of models that incorporate long- and short-term survival effects. We have extensively studies the properties of this estimation procedure by simulations (Year 2 report Section 8). In order for the models to be useful for the analysis of prostate cancer, we developed a hypothesis testing framework. We started with the traditional likelihood ratio test and variable selection procedures (Year 1 report, Section 6, Year 2 report Section 5), and then developed sophisticated techniques that allowed us to obtain exact observed profile information matrix for the models (Year 2 report Section 5). Coupled with the Wald test, this method made complex hypothesis testing computationally tractable. Subsequent sections of this report describe development of computer-intensive model selection and hypothesis testing procedures and application of the whole machinery to two large datasets on survival of prostate cancer patients.

4.2 Computer-intensive model selection

This section describes the work performed in Year 3.

Traditional forward and backward variable selection procedures as well as two computer intensive extended procedures (a Tree-Based procedure for the forward search and a Backwards Pooling procedure for the backwards search) were developed for all the models incorporated in the software.

Variable selection is organized by cycles of elementary hypotheses testing followed by a selection of the best model in the current cycle. The cycling is repeated until neither of the potential models evaluated at the current cycle satisfies a certain criteria for a continued search.

Technically, in our implementation of variable selection procedures, we always work with the maximal model containing all model parameters representing regression coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$. A non-cure model typically will not have an intercept term β_0. Forward variable selection procedures put maximal restrictions on the model parameters and work forward by removing those restrictions sequentially until either an unrestricted model emerges or a certain criterion is met. With the backwards selection procedures, an unrestricted maximal model is the starting point. The procedure then adds restrictions sequentially until either a maximally restricted model is achieved, or a certain stopping criterion is met.

4.2.1 Traditional variable selection procedures

With the forward selection procedure, a model where all model parameters representing regression coefficients $\beta = (\beta_1, \ldots, \beta_k)$ are fixed at zero. This corresponds to the hypothesis of homogeneity. The cycle of elementary hypothesis testing consists of evaluating a likelihood ratio characterizing the improvement in the goodness of fit resulting from removing a restriction for one of the regression coefficients $\beta_i = 0$. Model with the smallest p-value for testing the hypothesis represented by the restriction is chosen as the next base model, i.e. a model on which the next cycle will be based. The stopping occurs if all such p-values are larger than a certain threshold.

Forward procedures are subject to criticism that until the best model is achieved, hypothesis testing is based on a misspecified model, and therefore the validity of p-values may be a suspect. This consideration brings us to the forward procedures. Speed is an advantage of forward selection procedures as only models with degrees of freedom less than that of the best model are evaluated. Speed can further be improved by using a score test instead of likelihood ratio that requires fitting each potential model being evaluated (currently not implemented).

In the forward selection procedure an unrestricted maximal model is the starting point. Restrictions of the type $\mathrm{H}_0 : \beta_i = 0$ are evaluated at each point as the procedure cycles through all potential models in the current cycle where i goes through all yet unrestricted parameters. Model showing the largest p-value exceeding a certain threshold is selected as the base model for the next cycle. The procedure is stopped when non of such p-values exceed the threshold or when all parameters have been restricted. By the nature of this procedure, relatively big models with a sizable fraction of non-significant parameters are evaluated until the best one is achieved. Speed of the likelihood ratio test that requires fitting each potential model is a challenge. In order to improve on the speed of search, a Wald test have been implemented for all the models. The Wald test described in Section **??** uses our novel results on the exact profile information matrix (Section **??**). When occasional problems with information matrix inverses occur in the presence of many non-significant parameters (typically with highly overparameterized maximal models at the few initial dimension reduction steps), the procedure uses likelihood ratio test instead, and tries to swich back to the Wald test immediately after.

4.2.2 Backwards pooling procedure

In the backwards pooling procedure, two kinds of restrictions are considered:

- Fixing ith parameter $\mathrm{H}_0 : \beta_i = 0$, and
- Pooling ith and jth parameter $\mathrm{H}_0 : \beta_i = \beta_j$, $i \neq j$.

When the pooling hypothesis is first accepted, a cluster of two pooled parameters is created. Later on, any free parameter or a different cluster of pooled parameters could be merged with the first one. Note that by the nature of the imposed restrictions, any cluster of pooled parameters contributes one degree of freedom (one parameter-representative) to the model, and the pooling restriction forces all other mmembers of the cluster to be equal to the parameter-representative.

Unlike the traditional backwards variable selection procedure, its Backwards Pooling generalization is a much more flexible data mining tool. Indeed, the number of potential models considered on each cycle is an order of magnitude larger than in the traditional backwards selection procedure, and is equal to $A(A-1)/2+A$, where $A =$ (# of free parameters + # of pooled parameter clusters). To understand the difference, consider a model with single predictor $e^{\beta z}$ and one categorical covariate z. If simple contrast is used, z will be represented by a number of dummy variables comparing every possible category of z to a selected baseline category. Traditional variable selection in this situation would preserve any effect showing significant difference with the baseline. At the same time this "'best"' model may still be imperfect and overparameterised as some categories may show similar effects, and even though they might be significantly different from the baseline, they may show no difference whatsoever among themselves. The Backwards Pooling procedure will in this case continue variable selection until all differences between clusters of pooled parameters are significant. With z representing a categorized continuous veriable, for example, the procedure can be used to select optimal cutpoints on z that divide the sample into an optimal number of groups maximally seperated in terms of risk predicted by z. With factorial parameterization (the one that includes all possible interactions) of a set of categorical covariates, the output of the Backwards Pooling procedure is conceptually similar to a pruned regression tree, or to an output of a clustering algorithm, where groups correponding to distinct patterns of the covariate vector z are clustered so that the difference within each cluster in minimized, while the difference between clusters is maximized.

4.2.3 Tree-based methods

Traditional regression tree methodology [Breiman et al., 1984] is based on recursive partitioning of the data using splits defined by cutpoints put on covariates. The optimal cuntpoint at each step of the procedure typically maximizes a two-sample test statistic (minimizes the p-value). In survival analysis a logrank test or any other traditional rank test can be used to define the splits. We recognize, however, that logrank test is a score test for the Proportional Hazards model, and is sensitive to the long-term effect in the presence of long-term survivors. When long- and short-term effects are counteractive, logrank test can be dramatically underpowered. In Year 2 report (Figure 6, page 20; also see [Wendland et al., 2004]) we discussed a breast cancer example where due to counteracting long- and short-term effects survival curves crossed, and the logrank test showed a p-value of 0.79 while the test of homogeneity and separate tests for long- and short-term effects based on the PHPH model showed a highly significant result. Prostate cancer biochemical failure data from MSKCC also show crossing curves (Figure **??**). We therefore believe that in the presence of both long- and short-term effects, and particularly with crossing survival curves, PHPH or similar cure model with two predictors should be used as a basis for deriving the two-sample test. This will insure that regression tree would still be sensitive to differences departing from the proportional hazards assumption.

We have developed a general tree-based data mining procedure where splits are based on a two-sample test derived from any available Nonlinear Transformation Model (NTM).

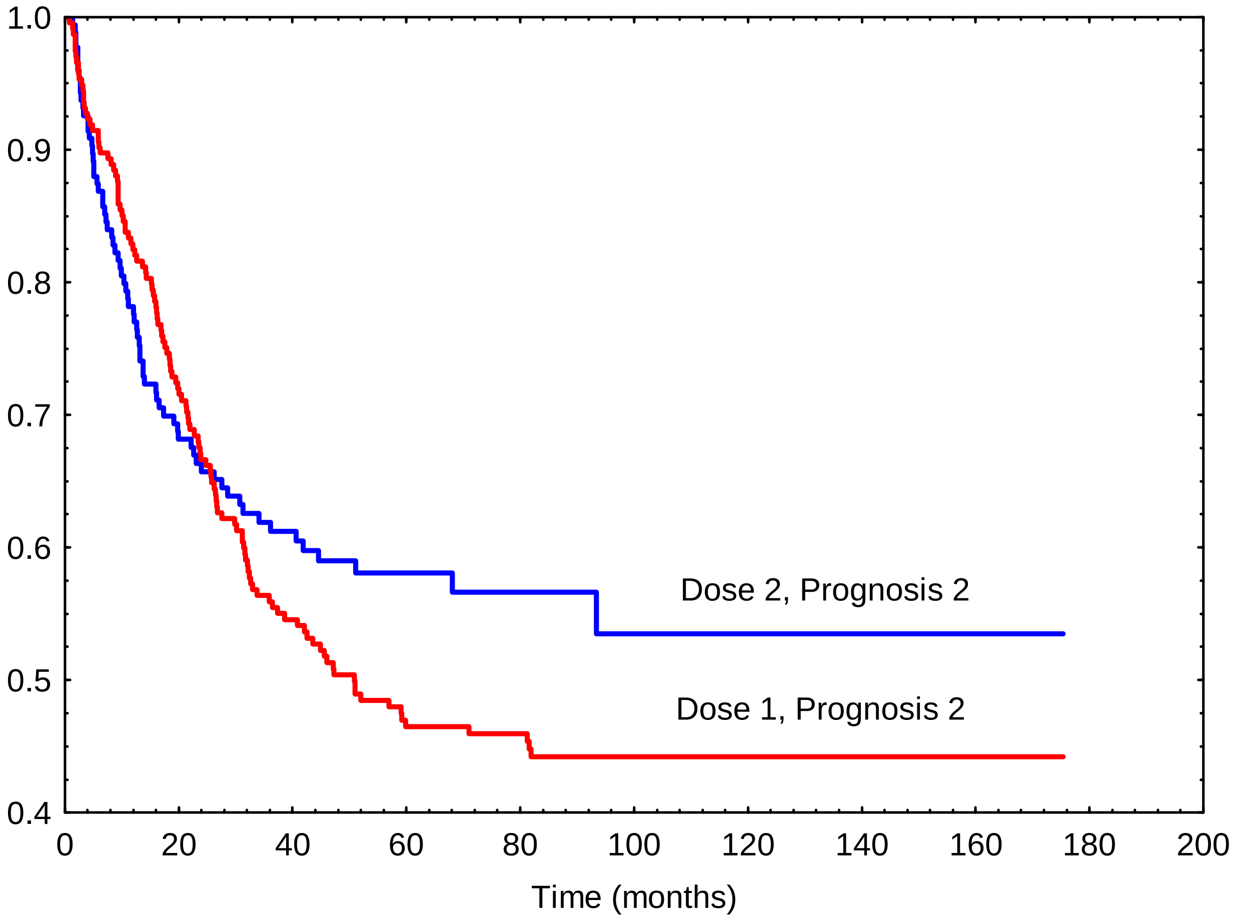


Figure 1: Time to biochemical failure. MSKCC data. Dose categories are 0 (lowest) through 3 (highest). Prognostic category is composed of PSA and Gleason score, categorized 0 (best) through 2 (worst).

Given a data set and an NTM the procedure does the following:

1. Recursively splits the data.
2. Finds the AIC of each node's model
3. Plots the tree with the AIC information

Splitting is organized according to the following algorithm.

1. For a continuous covariate x with distinct values $x_1, x_2, \ldots, x_k$, all possible splits of the form $x < x_i$ are considered. This is done via an indicator dummy variable that equals 1 if $x < x_i$, and 0 otherwise. Using a subset of the data corresponding to the node being considered for splitting, an NTM is fitted with the indicator covariate as the only variable in the model. A one step of the Powell optimization procedure is used with the profile likelihood as a target function. Profile likelihood corresponding to the outcome is retained as a criterion for the goodness of split.
2. For a discrete covariate x, with values $\{a_1, .., a_k\}$, all possible subsets of $\{a_1, ..., a_k\}$ are considered. For a given subset, the indicator dummy variable takes the value 1 if x is in the subset, and 0 otherwise. Evaluation of each such model is done as in the case of a continuous covariate.
3. Among all splits considered, the one with the largest profile likelihood is kept. This produces 2 descendant nodes where the procedure is repeated.
4. The stopping criterion is based on the minimal number of observations allowed in a node.

AIC is calculated using a full model fitting approach at each node. At a given node, using all the data, a model with covariates corresponding to the node in question and all its parent nodes is fitted. Node with the smallest AIC corresponds to the best model.

With a cure NTM with two predictors, same sets of covariates are used for long- and short term effects.

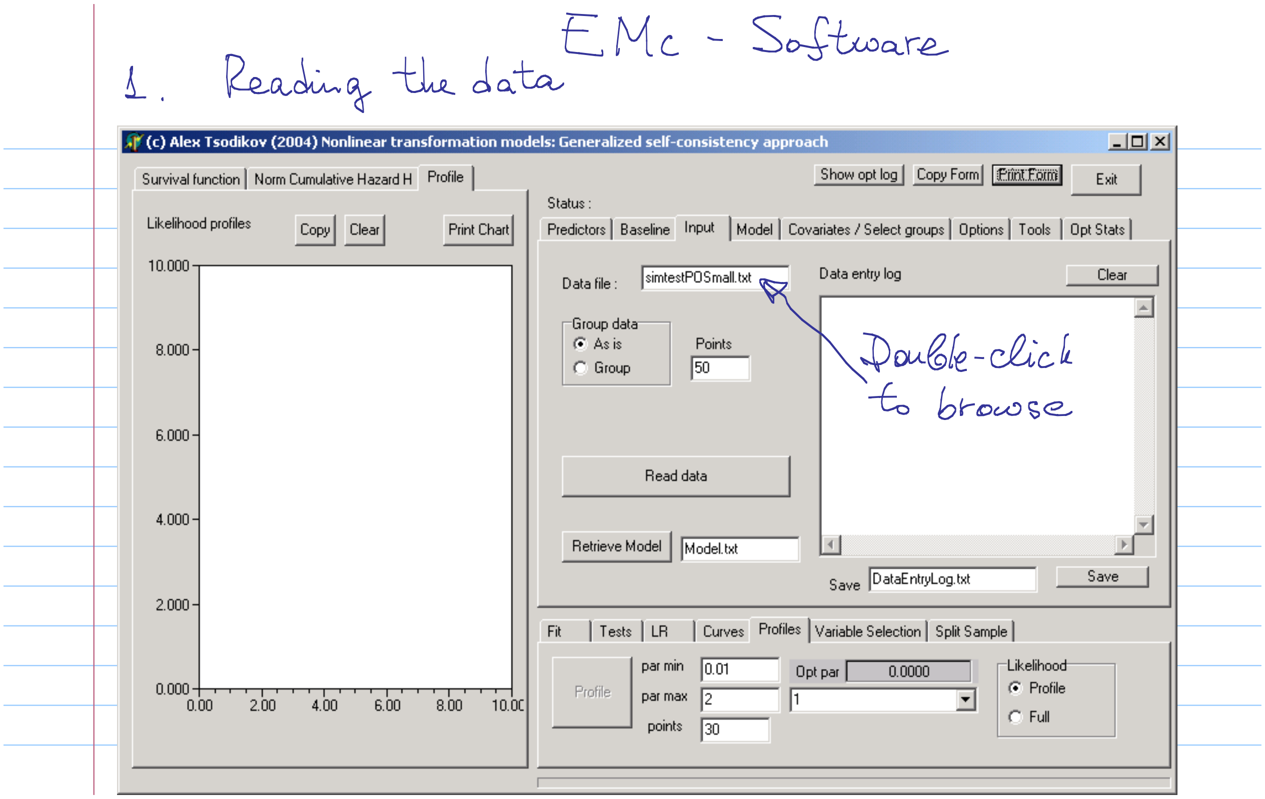
5 Computer Software and tools

This section describes the work performed in Year 2-3.

5.1 Delphi Package

The first version of the Delphi package EMc was described in our Year 2 report, Section 9. In Year 3 the package has seen many improvements and extensions. Key additions include development of the Backward Pooling variable selection procedure (Section 4.2.2) and incorporation of the profile information matrix theory (Section **??**) into hypothesis testing, confidence intervals and variable selection blocks. These additions helped reduce the computational burden of finding the best model for a large dataset such as MSKCC or SEER databases to the point that such analysis and data mining became feasible (Section 6). On the surface, these additions are seemless - when Backwards Pooling variable selction procedure is complete, the internal structure of the software is populated with the best model. Detailed log and result files are created.

The Figures 3-13 represent the basic fucntionality of the package.



1. Reading the data

EMc - Software

(c) Alex Tsodikov (2004) Nonlinear transformation models: Generalized self-consistency approach

Data file : simtestPOSmall.txt

Double-click to browse

Group data: ● As is ○ Group

Points: 50

Retrieve Model: Model.txt

Save: DataEntryLog.txt

par min: 0.01

par max: 2

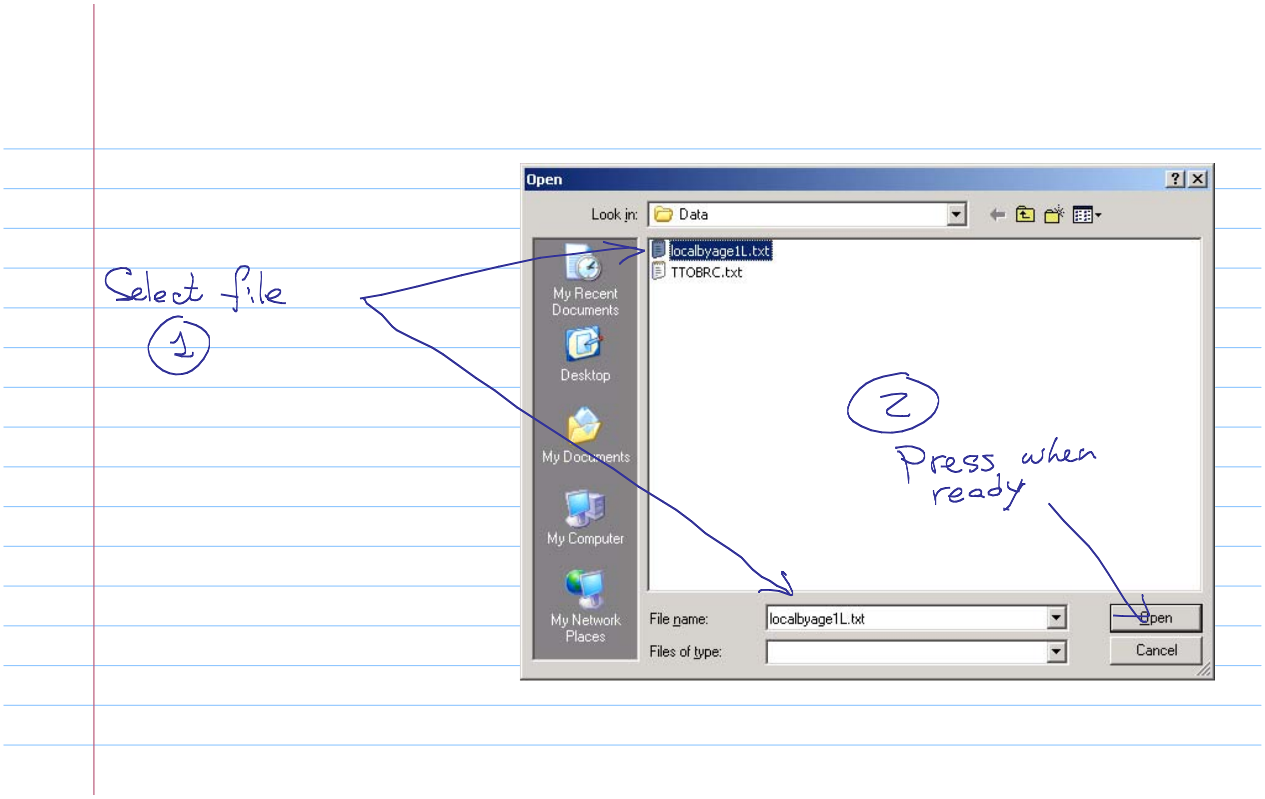
points: 30

Opt par: 0.0000

1

Likelihood: ● Profile ○ Full

Figure 2: Reading data into the Delphi EMc package.



Look in: Data

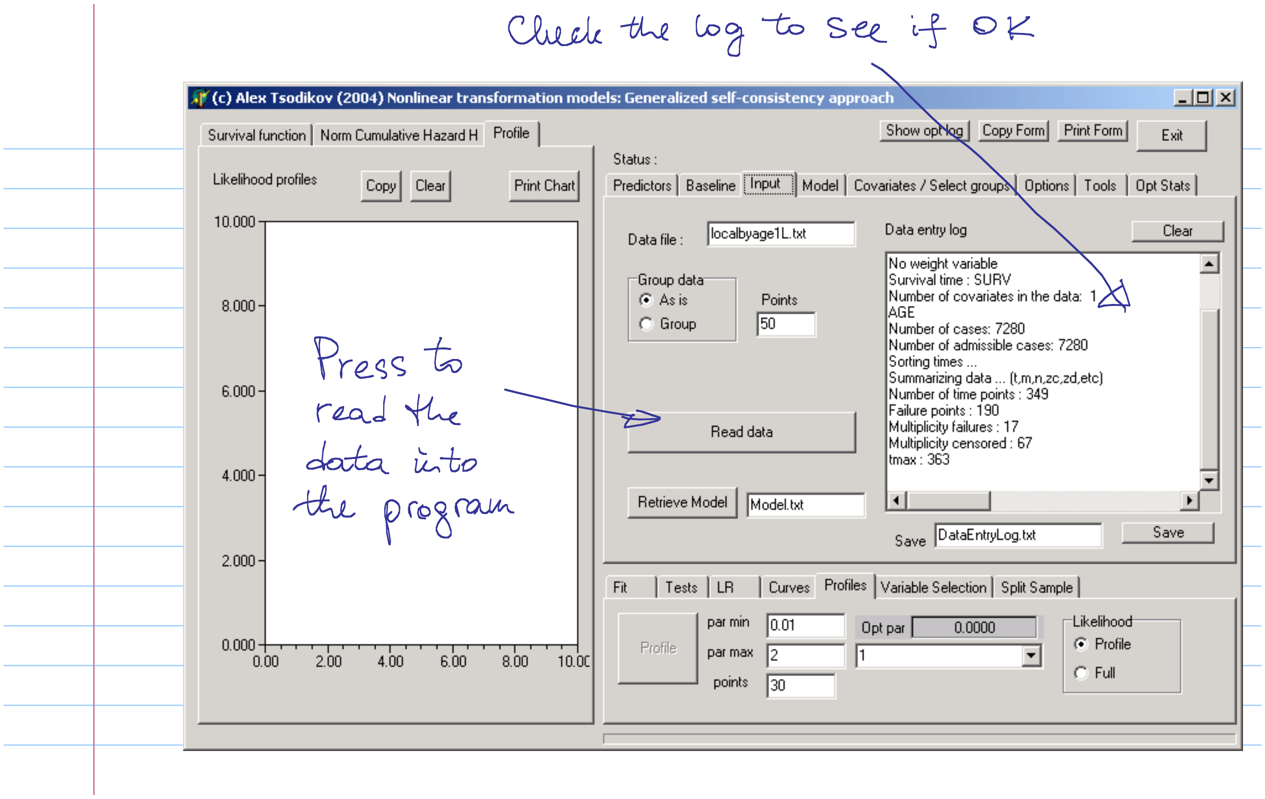
localbyage1L.txt

File name: localbyage1L.txt

Select file
1

2
Press when ready

Figure 3: Browsing for a data file.



Check the log to see if OK

Press to read the data into the program

(c) Alex Tsodikov (2004) Nonlinear transformation models: Generalized self-consistency approach

Data file : localbyage1L.txt

Group data: ● As is ○ Group

Points: 50

Retrieve Model: Model.txt

Data entry log

```
No weight variable
Survival time : SURV
Number of covariates in the data:  1
AGE
Number of cases: 7280
Number of admissible cases: 7280
Sorting times ...
Summarizing data ... (t,m,n,zc,zd,etc)
Number of time points : 349
Failure points : 190
Multiplicity failures : 17
Multiplicity censored : 67
tmax : 363
```

Save: DataEntryLog.txt

par min: 0.01

par max: 2

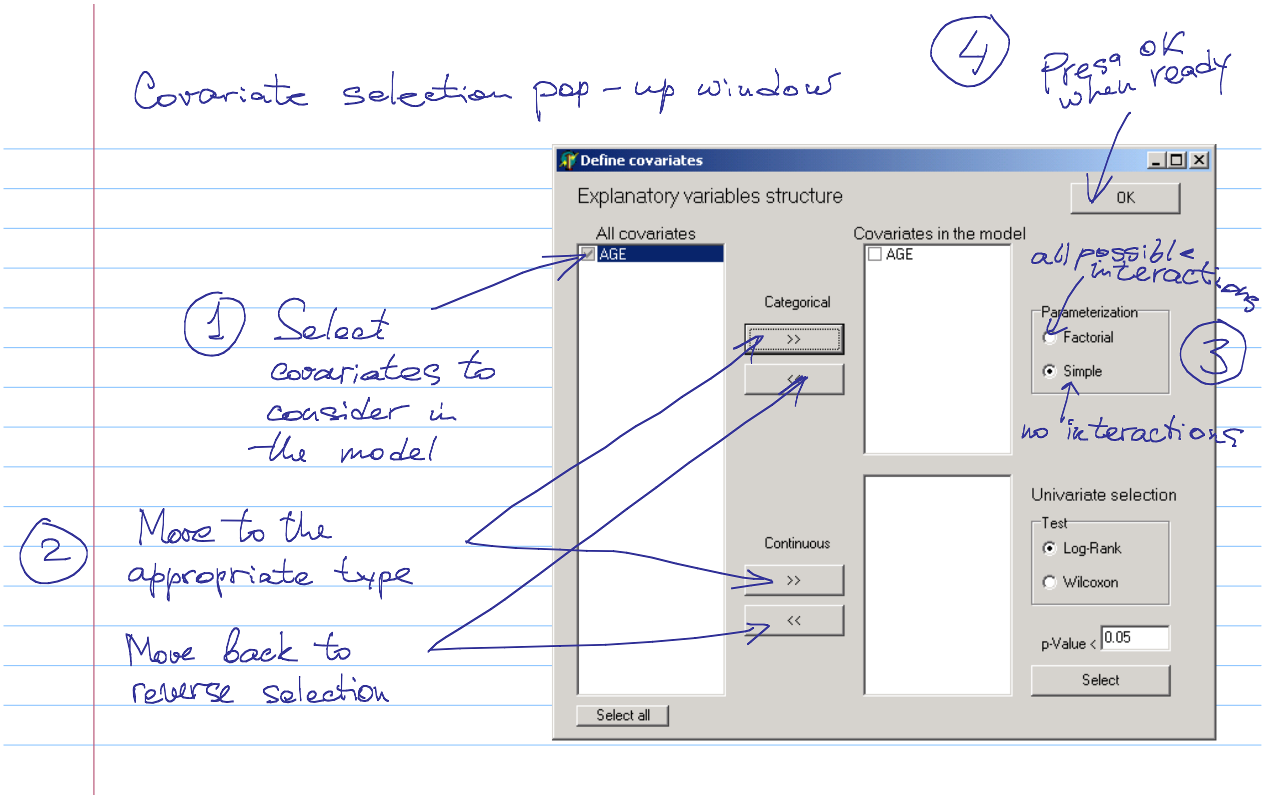
points: 30

Opt par: 0.0000

1

Likelihood: ● Profile ○ Full

Figure 4: Read the data and data entry log.



Explanatory variables structure

All covariates: ☑ AGE

Covariates in the model: □ AGE

Parameterization: ○ Factorial ● Simple

Univariate selection

Test: ● Log-Rank ○ Wilcoxon

p-Value < 0.05

Covariate selection pop-up window

4 Press OK when ready

1 Select covariates to consider in the model

2 Move to the appropriate type

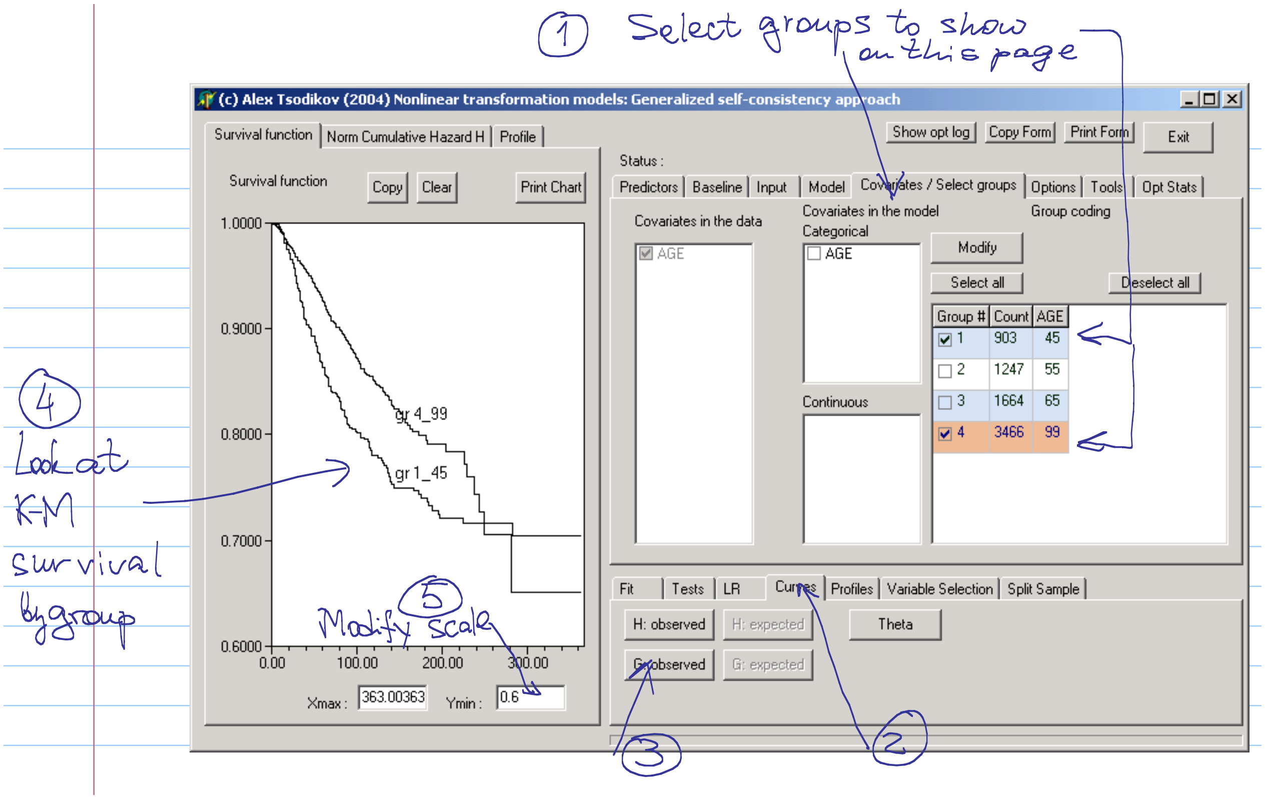
Move back to reverse selection

all possible interactions

3

no interactions

Figure 5: Select covariates to include in the maximal model and define their types.



☑ AGE (Covariates in the data)

Categorical: □ AGE

Group #	Count	AGE
☑ 1	903	45
□ 2	1247	55
□ 3	1664	65
☑ 4	3466	99

Xmax : 363.00363 Ymin : 0.6

1 Select groups to show on this page

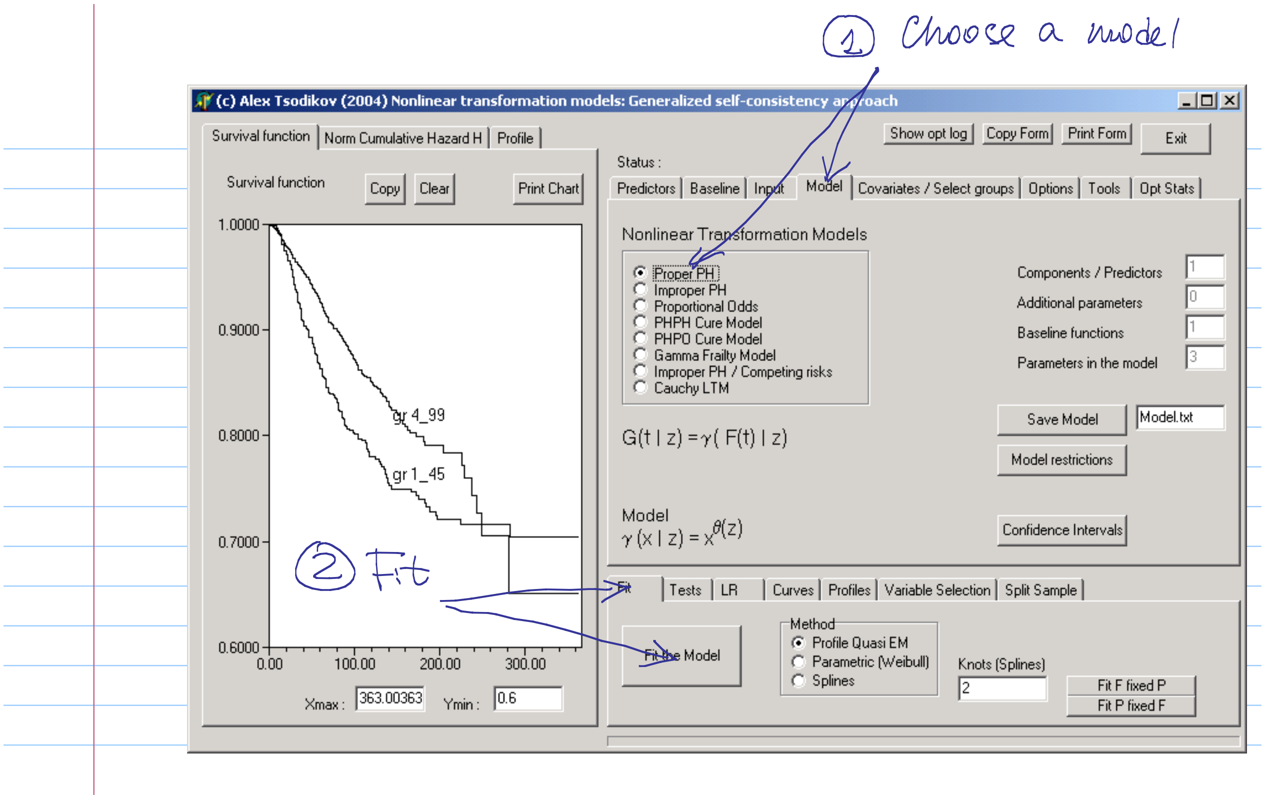
2

3

4 Look at K-M survival by group

5 Modify scale

Figure 6: Data slicing. Show survival curves corresponding to distinct patterns of categorical variables (groups).



1 Choose a model

2 Fit

(c) Alex Tsodikov (2004) Nonlinear transformation models: Generalized self-consistency approach

Nonlinear Transformation Models

- ● Proper PH
- ○ Improper PH
- ○ Proportional Odds
- ○ PHPH Cure Model
- ○ PHPO Cure Model
- ○ Gamma Frailty Model
- ○ Improper PH / Competing risks
- ○ Cauchy LTM

Components / Predictors	1
Additional parameters	0
Baseline functions	1
Parameters in the model	3

$G(t \mid z) = \gamma(F(t) \mid z)$

Model

$\gamma(x \mid z) = x^{\theta(z)}$

Method

- ● Profile Quasi EM
- ○ Parametric (Weibull)
- ○ Splines

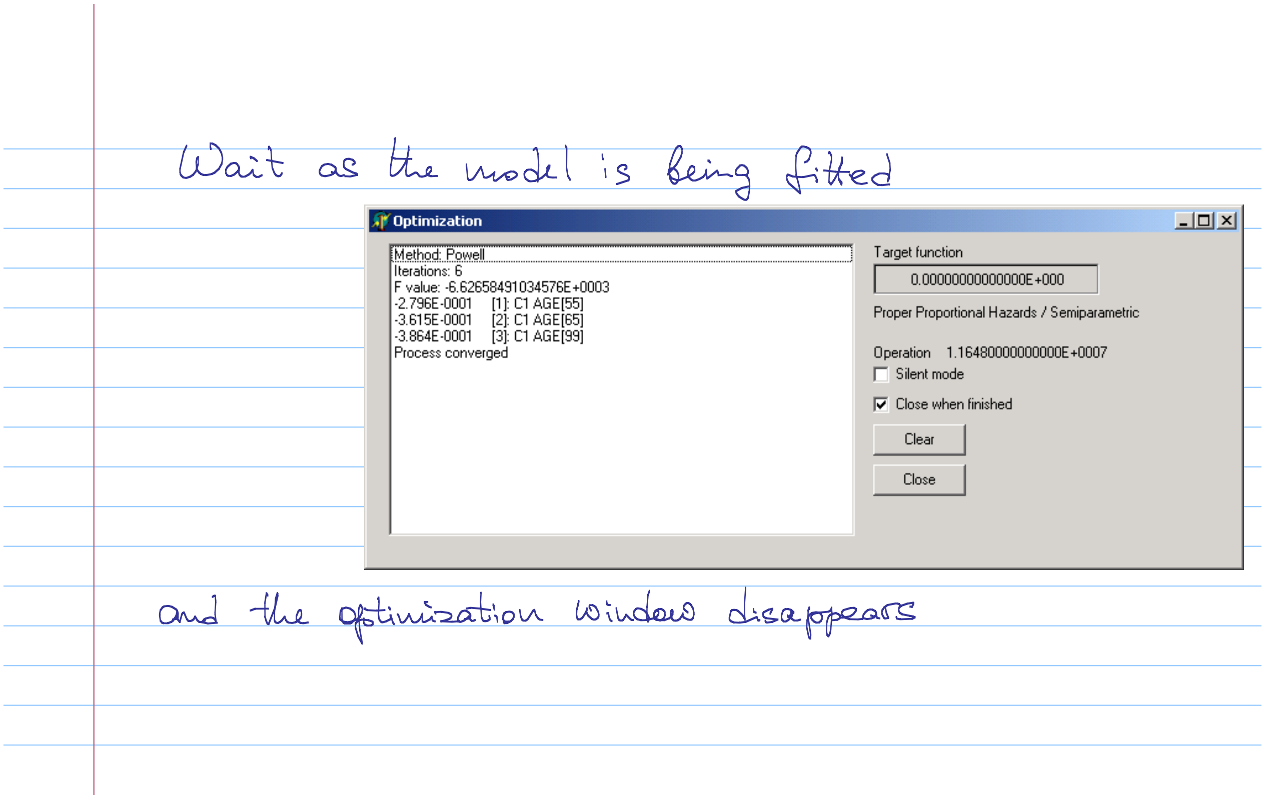
Knots (Splines): 2

Xmax : 363.00363

Ymin : 0.6

Model.txt

Figure 7: Choose a model to be fitted.



Wait as the model is being fitted

```
Method: Powell
Iterations: 6
F value: -6.62658491034576E+0003
-2.796E-0001    [1]: C1 AGE[55]
-3.615E-0001    [2]: C1 AGE[65]
-3.864E-0001    [3]: C1 AGE[99]
Process converged
```

Target function: 0.00000000000000E+000

Proper Proportional Hazards / Semiparametric

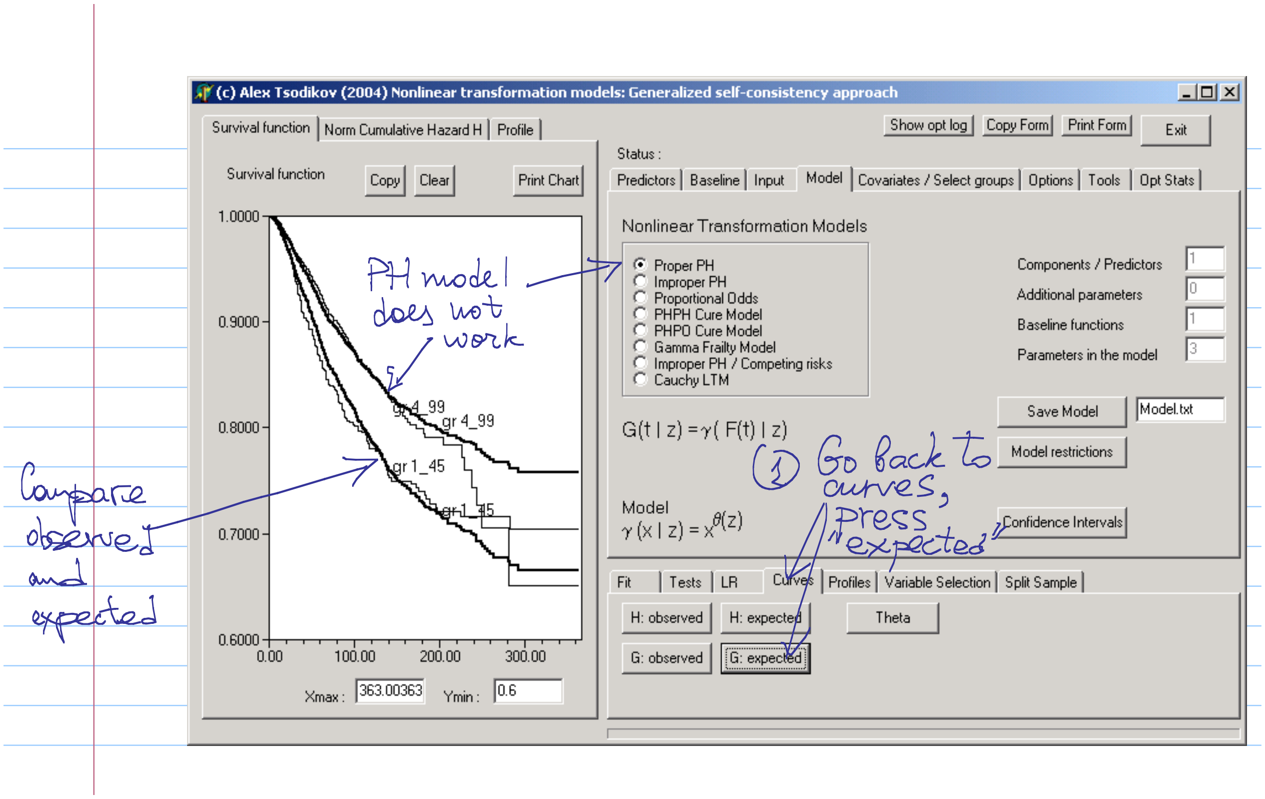
Operation 1.16480000000000E+0007

□ Silent mode

☑ Close when finished

and the optimization window disappears

Figure 8: Interim output of the likelihood maximization procedure.



● Proper PH
○ Improper PH
○ Proportional Odds
○ PHPH Cure Model
○ PHPO Cure Model
○ Gamma Frailty Model
○ Improper PH / Competing risks
○ Cauchy LTM

Components / Predictors: 1
Additional parameters: 0
Baseline functions: 1
Parameters in the model: 3

Model.txt

$G(t \mid z) = \gamma(F(t) \mid z)$

Model
$\gamma(x \mid z) = x^{\theta(z)}$

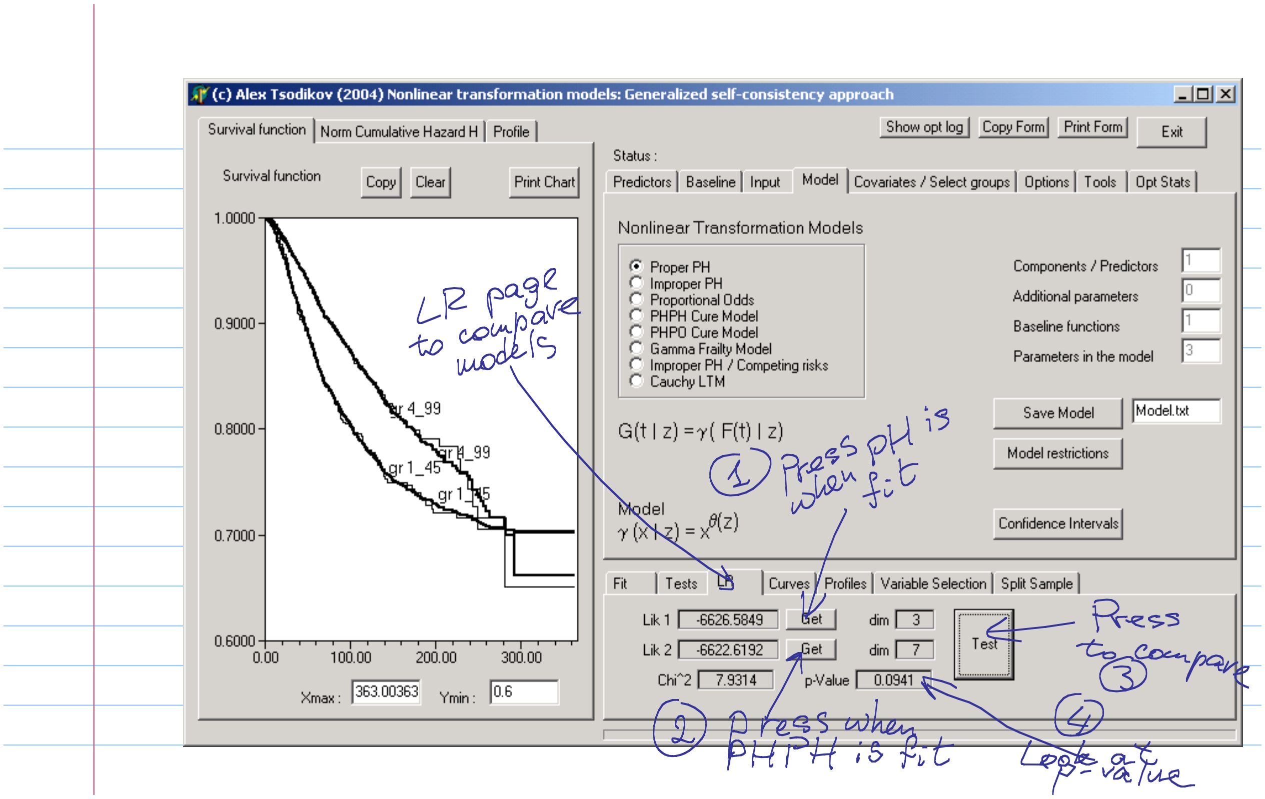
Xmax : 363.00363
Ymin : 0.6

PH model does not work

Compare observed and expected

(1) Go back to curves, press "expected"

Figure 9: Superimposing survival curves expected under the model on observed Kaplan-Meier estimates for distinct patterns of categorical variables.



Xmax : 363.00363 Ymin : 0.6

● Proper PH
○ Improper PH
○ Proportional Odds
○ PHPH Cure Model
○ PHPO Cure Model
○ Gamma Frailty Model
○ Improper PH / Competing risks
○ Cauchy LTM

Components / Predictors 1
Additional parameters 0
Baseline functions 1
Parameters in the model 3

Model.txt

$G(t \mid z) = \gamma(F(t) \mid z)$

Model
$\gamma(x \mid z) = x^{\theta(z)}$

Lik 1 -6626.5849 dim 3
Lik 2 -6622.6192 dim 7
Chi^2 7.9314 p-Value 0.0941

LR page to compare models

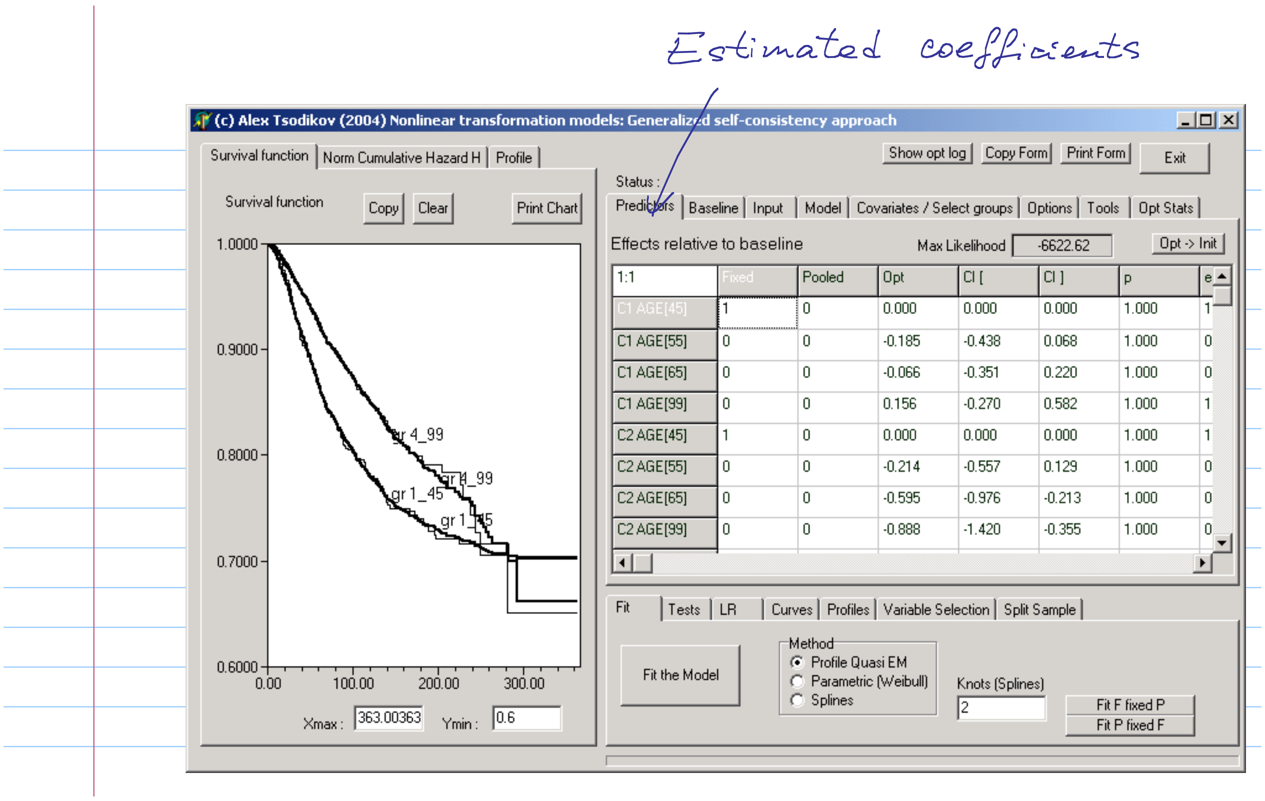
① Press when PH is fit

② Press when PHPH is fit

Press to compare ③

④ Look at p-value

Figure 10: Performing a likelihood ratio test to compare nested models.



Estimated coefficients

(c) Alex Tsodikov (2004) Nonlinear transformation models: Generalized self-consistency approach

Xmax : 363.00363 Ymin : 0.6

Effects relative to baseline

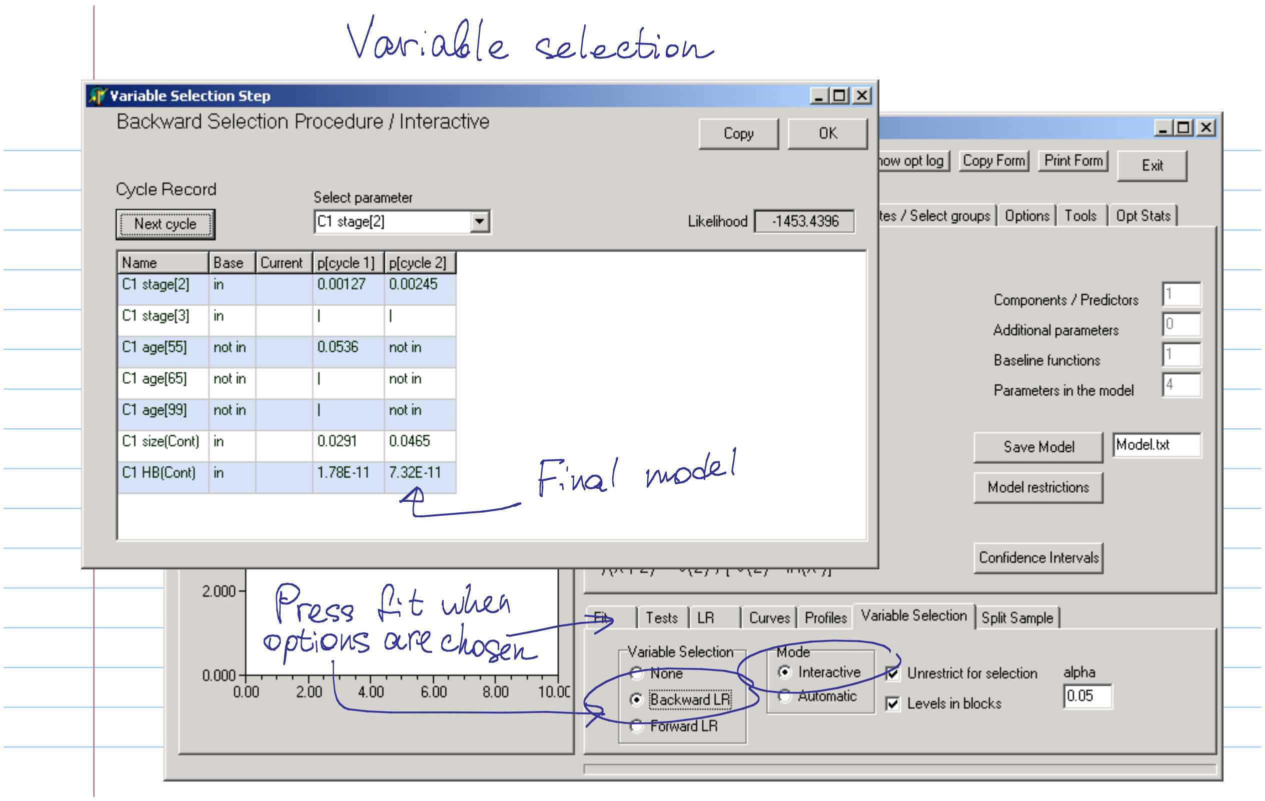
Max Likelihood -6622.62

1:1	Fixed	Pooled	Opt	CI [	CI]	p	e
C1 AGE[45]	1	0	0.000	0.000	0.000	1.000	1
C1 AGE[55]	0	0	-0.185	-0.438	0.068	1.000	0
C1 AGE[65]	0	0	-0.066	-0.351	0.220	1.000	0
C1 AGE[99]	0	0	0.156	-0.270	0.582	1.000	1
C2 AGE[45]	1	0	0.000	0.000	0.000	1.000	1
C2 AGE[55]	0	0	-0.214	-0.557	0.129	1.000	0
C2 AGE[65]	0	0	-0.595	-0.976	-0.213	1.000	0
C2 AGE[99]	0	0	-0.888	-1.420	-0.355	1.000	0

Method
● Profile Quasi EM
○ Parametric (Weibull)
○ Splines

Knots (Splines) 2

Figure 11: Point estimates and confidence intervals for fitted model parameters.



Variable selection

Variable Selection Step

Backward Selection Procedure / Interactive

Cycle Record

Select parameter: C1 stage[2]

Likelihood: -1453.4396

Name	Base	Current	p[cycle 1]	p[cycle 2]
C1 stage[2]	in		0.00127	0.00245
C1 stage[3]	in		|	|
C1 age[55]	not in		0.0536	not in
C1 age[65]	not in		|	not in
C1 age[99]	not in		|	not in
C1 size(Cont)	in		0.0291	0.0465
C1 HB(Cont)	in		1.78E-11	7.32E-11

Final model

Components / Predictors: 1
Additional parameters: 0
Baseline functions: 1
Parameters in the model: 4

Save Model: Model.txt

Variable Selection
○ None
● Backward LR
○ Forward LR

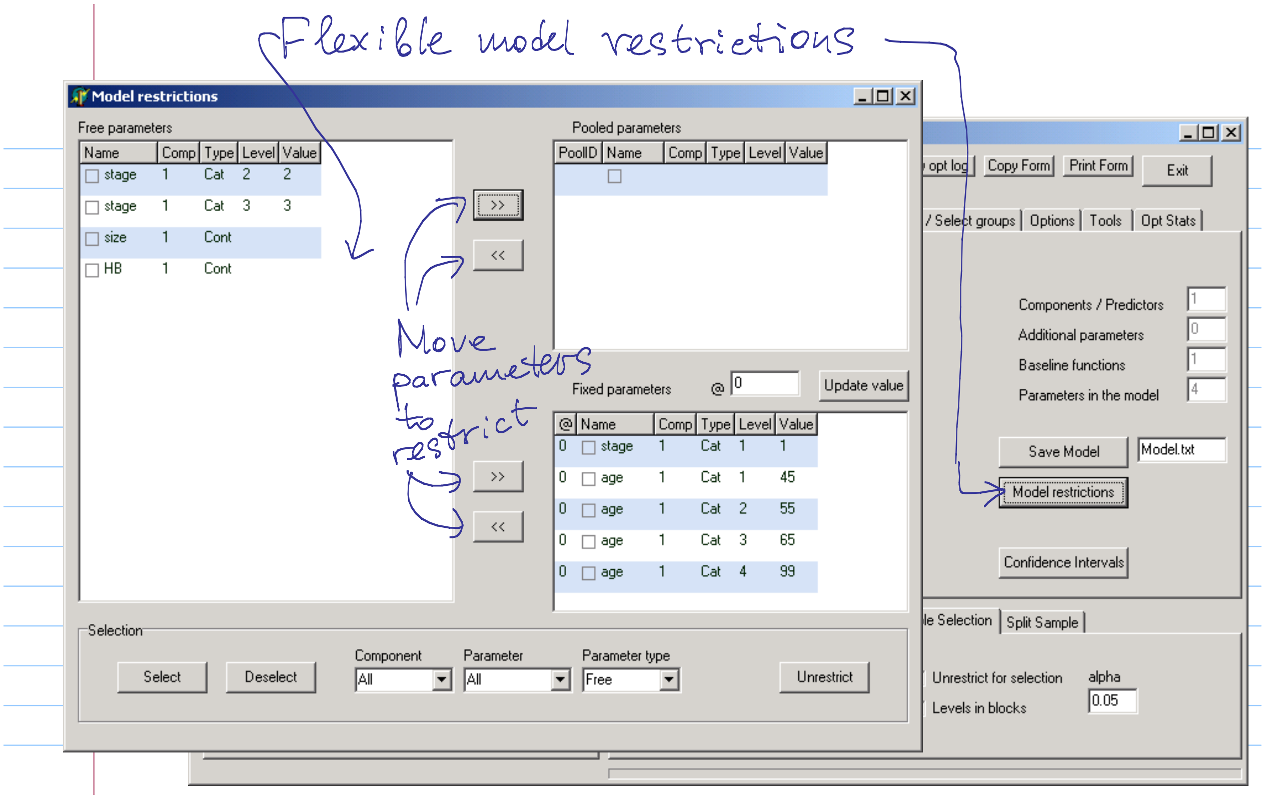
Mode
● Interactive
○ Automatic

☑ Unrestrict for selection
☑ Levels in blocks

alpha: 0.05

Press fit when options are chosen

Figure 12: Variable selection procedures and interim output.


Flexible model restrictions

Move parameters to restrict

Model restrictions

Free parameters

Name	Comp	Type	Level	Value
□ stage	1	Cat	2	2
□ stage	1	Cat	3	3
□ size	1	Cont		
□ HB	1	Cont		

Pooled parameters

PoolID	Name	Comp	Type	Level	Value
	□				

Fixed parameters @ 0

@	Name	Comp	Type	Level	Value
0	□ stage	1	Cat	1	1
0	□ age	1	Cat	1	45
0	□ age	1	Cat	2	55
0	□ age	1	Cat	3	65
0	□ age	1	Cat	4	99

Selection

Component: All

Parameter: All

Parameter type: Free

Components / Predictors: 1

Additional parameters: 0

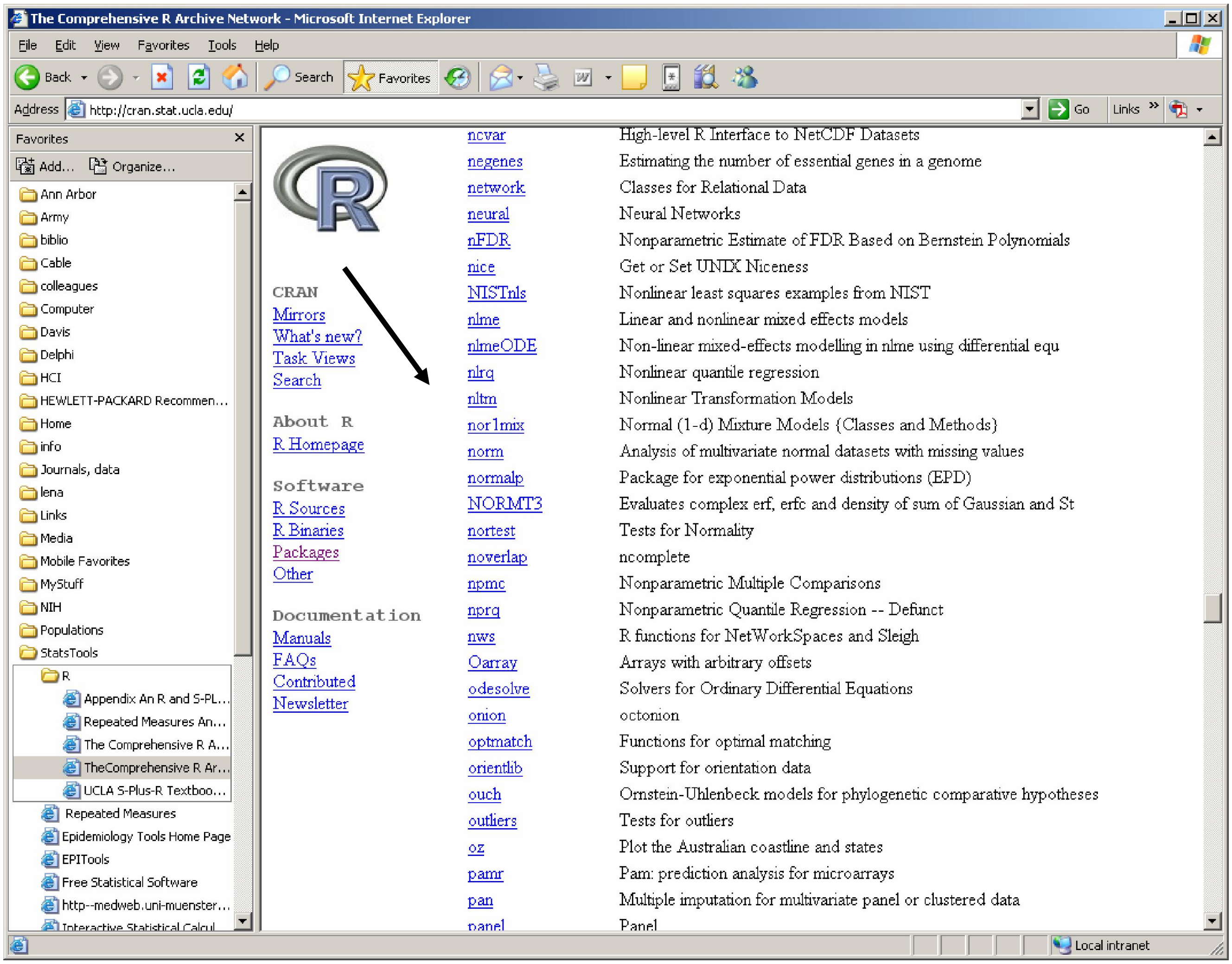
Baseline functions: 1

Parameters in the model: 4

Save Model: Model.txt

alpha: 0.05

Figure 13: Flexible model restrictions. Manually specifying a restricted model to test specific hypotheses.



ncvar	High-level R Interface to NetCDF Datasets
negenes	Estimating the number of essential genes in a genome
network	Classes for Relational Data
neural	Neural Networks
nFDR	Nonparametric Estimate of FDR Based on Bernstein Polynomials
nice	Get or Set UNIX Niceness
NISTnls	Nonlinear least squares examples from NIST
nlme	Linear and nonlinear mixed effects models
nlmeODE	Non-linear mixed-effects modelling in nlme using differential equ
nlrq	Nonlinear quantile regression
nltm	Nonlinear Transformation Models
nor1mix	Normal (1-d) Mixture Models {Classes and Methods}
norm	Analysis of multivariate normal datasets with missing values
normalp	Package for exponential power distributions (EPD)
NORMT3	Evaluates complex erf, erfc and density of sum of Gaussian and St
nortest	Tests for Normality
noverlap	ncomplete
npmc	Nonparametric Multiple Comparisons
nprq	Nonparametric Quantile Regression -- Defunct
nws	R functions for NetWorkSpaces and Sleigh
Oarray	Arrays with arbitrary offsets
odesolve	Solvers for Ordinary Differential Equations
onion	octonion
optmatch	Functions for optimal matching
orientlib	Support for orientation data
ouch	Ornstein-Uhlenbeck models for phylogenetic comparative hypotheses
outliers	Tests for outliers
oz	Plot the Australian coastline and states
pamr	Pam: prediction analysis for microarrays
pan	Multiple imputation for multivariate panel or clustered data
panel	Panel

Figure 14: Link to nltm package on the R project web page.

5.2 R package nltm

The nltm package implements the basic functionality of EMc (Section 5.1) in the R environment without the Delphi GUI. Although these two packages are implementations of the same basic methodology, the source code is written in two different languages. EMc is written in Delphi 7 by Inprise, a visual and object oriented version of the Pascal programming language. The R package nltm is written in c and R with the subsequent compilation as an R package. The source codes and a brief package description is available on the R webpage (Figure 14). Compiled versions for Unix and Windows will be produced by the R project team from the source codes, and at this point the package will be directly installable from the user's R environment. In the meanwhile, both packages compiled for Windows are available from the PI for a manual installation. The following models are currently supported by nltm:

- Proportional hazard model (PH):
$$G(t|z) = F(t)^{\theta(z)}$$
- Proportional hazard cure model (PHC):
$$G(t|z) = \exp(-\theta(z)(1 - F(t)))$$

- Proportional odds model (PO):
$$G(t|z) = \frac{\theta(z)}{\theta(z) - \log F(t)}$$

- Proportional hazard - proportional hazard cure model (PHPHC):
$$G(t|z) = \exp(-\theta(z)(1 - F^{\eta(z)}(t)))$$

- Proportional hazard - proportional odds cure model (PHPOC):
$$G(t|z) = \exp\left\{-\theta(z)\frac{1 - F(t)}{1 - (1 - \eta(z))F(t)}\right\}$$

- Gamma frailty model (GFM):
$$G(t|z) = \left[\frac{\theta(z)^{\eta(z)}}{\theta(z) - \ln(F(t))}\right]^{\eta(z)}$$

- Proportional hazard - proportional odds model (PHPO):
$$G(t|z) = \frac{\theta F^{\eta(z)}(t)}{1 - (1 - \theta)F^{\eta(z)}(t)}$$

The following R command represents the syntax for a call to an estimation procedure

```
nltm(formula=formula(data), data=parent.frame(), subset, na.action,
init, control, model=c("PH","PHC","PO","PHPHC","PHPOC","GFM","PHPO"),
verbose=FALSE, ...),
```

where the arguments have the following meaning

`formula` A formula object, with the response on the left of a ”‘ ”’ operator, and the terms on the right. The response must be a survival object as returned by the `Surv` function.

`data` A `data.frame` in which to interpret the variables named in the `formula`, or in the `subset` argument.

`subset` Expression saying that only a subset of the rows of the data should be used in the fit.

`na.action` A missing-data filter function, applied to the `model.frame`, after any subset argument has been used. Default is `options()$na.action`.

`init` Vector of initial values for the calculation of the maximum likelihood estimator of the regression parameters. Default initial value is zero.

`control` Object of class `coxph.control` specifying iteration limit and other control options. Default is `nltm.control(...)`.

`model` A character string specifying a non-linear transformation model. Default Proportional Hazards Model.

`verbose` If `TRUE` it stores information from maximization of likelihood and calculation of information matrix in a file. Default is `FALSE`.

... Other arguments

The procedure returns a value (object) of the class `"‘coxph"’`. Thus, in its usage the package is similar to existing R package `survival`.

The following is an example of usage in R.

1. Create a simple test data set with four variables, `time` (survival time), `status` (censoring index), and two covariates, `age` (categorized age in years represented by the right point of the bin), and `size` (tumor size in mm).

   ```
   test1 <- list(time=c(10,7,32,23,22,6,16,34,32,25,11,20,19,6,17,35,6,13,9,6,1),
   status=c(1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0),
   size=c(1.79,7.93,2.02,6.89,2.30,7.82,1.25,9.85,6.02,3.43,4.72,7.45,8.83,
    9.53,1.10,1.06,5.25,5.86,2.03,3.62,3.52),
   age=factor(c(65,65,65,65,99,45,65,99,99,99,65,45,65,55,45,45,55,55,55,99,65)))
   ```

2. Call ntml procedure to fit the PO model

   ```
   nltm(Surv(time,status) ~ size + age, data=test1, model="PO")
   ```

   results in a table of point estimates, standard errors and p-values for dropping the term.

   ```
   Call:
   nltm(formula = Surv(time, status) ~ size + age, data = test1,
       model = "PHPHC")


           coef exp(coef) se(coef)      z     p
   size   0.141  1.15e+00    0.174  0.812 0.420
   age55  5.000  1.48e+02    9.763  0.512 0.610
   age65  1.736  5.67e+00    1.219  1.424 0.150
   age99 -0.264  7.68e-01    1.450 -0.182 0.860
   size  -0.117  8.90e-01    0.234 -0.500 0.620
   age55 -4.975  6.91e-03    9.815 -0.507 0.610
   age65 -3.353  3.50e-02    1.796 -1.866 0.062
   age99 -3.358  3.48e-02    2.251 -1.492 0.140
   cure  -1.920  1.47e-01    1.368 -1.403 0.160

   Likelihood ratio test=10.3  on 9 df, p=0.325  n= 21
   ```

3. A similar call to the PHPH cure model with long- and short-term predictors results in a table twice the size of the one for a PO model, where the first set of coefficients corresponds to long-term effects of the covariate, and the second set of coefficients corresponds to short-term effects of the same covariates.

```
nltm(Surv(time,status) ~ size + age, data=test1, model="PHPHC")

Call:
nltm(formula = Surv(time, status) ~ size + age, data = test1,
    model = "PHPHC")


        coef exp(coef) se(coef)      z     p
size   0.141  1.15e+00    0.174  0.812 0.420
age55  5.000  1.48e+02    9.763  0.512 0.610
age65  1.736  5.67e+00    1.219  1.424 0.150
age99 -0.264  7.68e-01    1.450 -0.182 0.860
size  -0.117  8.90e-01    0.234 -0.500 0.620
age55 -4.975  6.91e-03    9.815 -0.507 0.610
age65 -3.353  3.50e-02    1.796 -1.866 0.062
age99 -3.358  3.48e-02    2.251 -1.492 0.140
cure  -1.920  1.47e-01    1.368 -1.403 0.160

Likelihood ratio test=10.3  on 9 df, p=0.325  n= 21
```

5.3 R package rpNLTM

The R package `rpNLTM` implements Recursive Partitioning and Regression Trees algorithms using splits with criteria supplied by fitting an NTM model using `nltm` package functions.

The following R command represents the syntax for a call to an estimation procedure

```
rpNLTM(formula, data, weights, subset, na.action=na.rpart, method,
model=FALSE, x=FALSE, y=TRUE, parms, control, cost, control.nltm,
model.nltm=c("PH","PHC","PO","PHPHC","PHPOC","GFM","PHPO"), verbose=FALSE, ...)
```

where the arguments have the following meaning

`formula` A formula, as in the `lm` function.

`data` An optional data frame in which to interpret the variables named in the formula weights optional case weights.

`subset` Optional expression saying that only a subset of the rows of the data should be used in the fit.

`na.action` The default action deletes all observations for which y is missing, but keeps those in which one or more predictors are missing.

`method` One of ”`anova`”, ”`poisson`”, ”`class`”, ”`exp`” or ”`nltm`”. ”`nltm`” method is of primary interest as it triggers regression trees procedures based on the methodology developed in this project. The other methods are inherited from the existing package ”`rpart`” and are not related to this project.

`model` If logical: keep a copy of the model frame in the result? If the input value for model is a model frame (likely from an earlier call to the rpart function), then this frame is used rather than constructing new data.

`x` Keep a copy of the x matrix in the result.

`y` Keep a copy of the dependent variable in the result. If missing and model is supplied this defaults to FALSE.

`parms` Optional parameters for the splitting function. Anova splitting has no parameters. Poisson splitting has a single parameter, the coefficient of variation of the prior distribution on the rates. The default value is 1. Exponential splitting has the same parameter as Poisson. For classification splitting, the list can contain any of: the vector of prior probabilities (component prior), the loss matrix (component loss) or the splitting index (component split). The priors must be positive and sum to 1. The loss matrix must have zeros on the diagnoal and positive off-diagonal elements. The splitting index can be gini or information. The default priors are proportional to the data counts, the losses default to 1, and the split defaults to gini.

`control` Options that control details of the rpart algorithm.

`cost` A vector of non-negative costs, one for each variable in the model. Defaults to one for all variables. These are scalings to be applied when considering splits, so the improvement on splitting on a variable is divided by its cost in deciding which split to choose.

`control.nltm` Object of class `coxph.control` specifying iteration limit and other control options. Default is `nltm.control(...)`.

`model.nltm` A character string specifying a non-linear transformation model. Default Proportional Hazards Model.

`verbose` If TRUE it stores information from maximization of likelihood and calculation of information matrix in a file. Default is FALSE.

`...` arguments to `rpart.control` may also be specified in the call to `rpart`. They are checked against the list of valid arguments.

The procedure returns an object of class `rpart`, a superset of class `tree`.
The following is an example of usage.

1. Introduce a dataset.

```
leuk1 <- list(time=c(10,7,32,23,22,6,16,34,32,25,11,20,19,6,17,35,6,13,9,6,10),
status=c(1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0),
stage=factor(c(2, 1, 0, 0, 1, 2, 1, 1, 2, 0, 0, 1, 0, 0, 2, 0, 2, 1, 1, 0, 0)),
size=c(1.79,7.93,2.02,6.89,2.30,7.82,1.25,9.85,6.02,3.43,4.72,7.45,8.83,
9.53,1.10,1.06,5.25,5.86,2.03,3.62,3.52),
age=factor(c(65,65,65,65,99,45,65,99,99,99,65,45,65,55,45,45,55,55,55,99,65)))
```

2. Generate a regression tree based on the PO model.

```
fit <- rpNLTM(formula=Surv(time,status) ~ size + age, data=leuk1,
method="nltm", model.nltm="PO", verbose=TRUE, minsplit=5)
```

3. Plot the regression tree.

   ```
   plot(fit)
   ```

4. Mark te nodes of the tree.

   ```
   text(fit, pretty=TRUE, all=TRUE)
   ```

   The result is shown on Figure 15.

6 Data Analysis

This section describes the work performed in Year 3.

6.1 Memorial Sloan Kettering Cancer Center Database

We have conducted data analysis using four different endpoints.

Biochemical This endpoint is defined by ASTRO as three successive prostate specific antigen (PSA) elevations observed from a post-treatment nadir PSAlevel. This endpoint is also termed "‘PSA relapse"’ or "‘PSA failure"’ (IJROBP 37:1035-1041 1997).

Local failure Local failure is defined as palpable recurrence and/or positive re-biopsy (Biopsy-confirmed recurrence).

Distant Distant failure represents detection of distant metastasis (DM).

Survival This failure id defined as prostate cancer-specific death (cause-specific failure).

Other than local failure, all end points examined may stem from two sources: subclinical metastases already present at the time of treatment or shedding of tumor cells that were not sterilized during radiation therapy.

The significance of models accomodating long- and short-term effects developed in this project is that they allow us to reproduce complex the timing patterns of failures resulting from failures originating from different biological sources.

The long-term effect represents a combination of long-term patient's prognosis based on clinical characteristics of the disease at diagnosis, and the treatment effect representing the chance that we eradicate tumor cells at the time of treatment. It is therefore expected that with the prognostic index being fixed, a more radical treatment would result in a higher cure rate.

At the same time treatment may affect metastatic cells is a different way or exert a "‘survival of the fittest"’ effect on the residual tumor cells giving rise to a failure. These effects have the potential of altering the dynamics and timing of the development of failure without necessarily affecting the long-term survival determined by the probability that no residual tumor cells are present rather than by their biological characteristics.

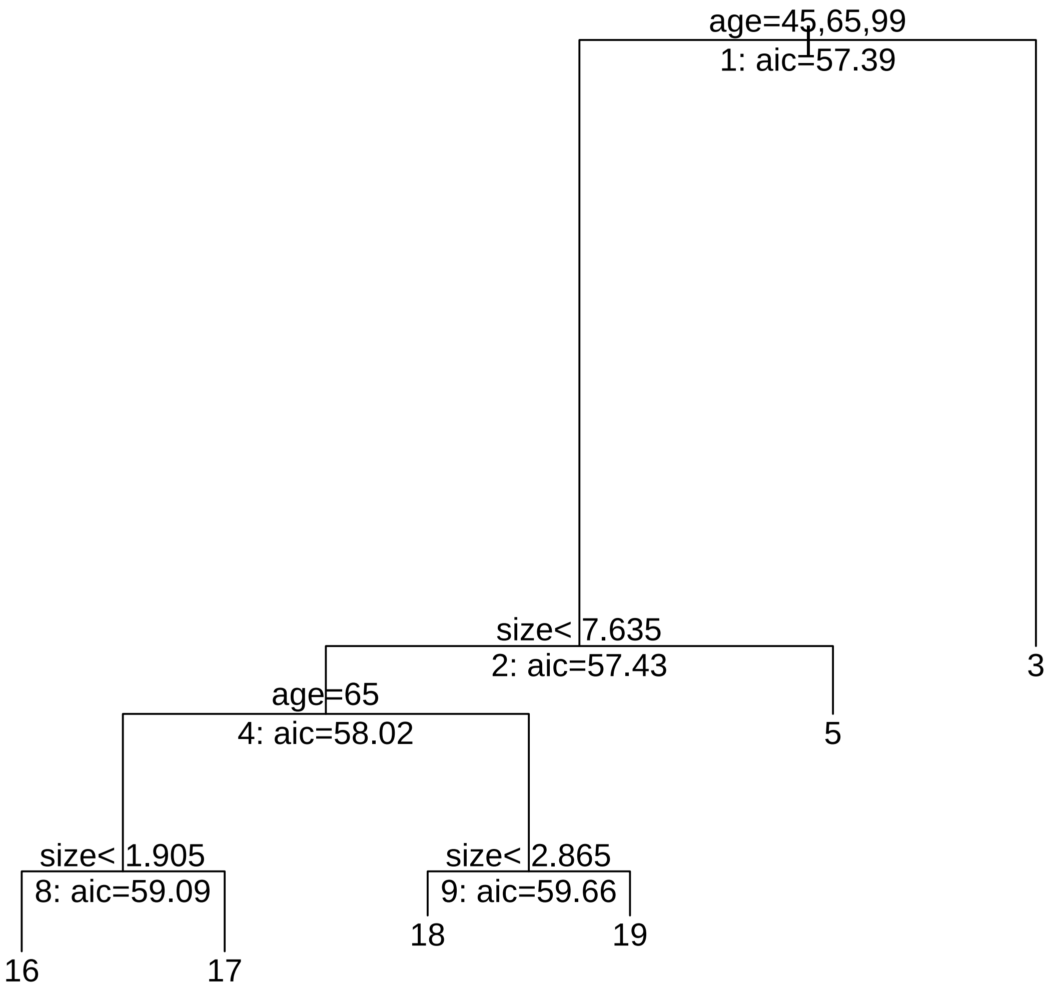


Figure 15: An example of the regression tree built using a PO nonlinear transformation model.

Having the biostatistical instrument at our disposal that recognizes the above mentioned complexity, we approach the analysis using a nonlinear transformation cure models with two predictors.

In our preliminary data analysis reported in Year 2 progress report (Section 8) we have identified two models potentially suitable for complex survival responses observed in prostate cancer: the Gamma Frailty model with covariates incorporated into its scale and shape parameter, and the so-called PHPH cure model. We also observed that PHPH model was the only one that allowed us to reproduce crossing survival curves. Having obtained an extended version of the MSKCC database of patients undergoing radiation therapy for prostate cancer, we have observed crossing survival curves for some of the covariate groups as shown, for example, in Figure 1. Based on this observation, we decided to select the PHPH model as our primary tool of data analysis. Our prior experience suggests, however, that other cure models with two predictors such as PHPOC, or a two component mixture model, would give very similar results.

The population examined in this study consists of 1765 patients with biopsy-confirmed localized prostate cancer who were treated at MSKCC with external-beam radiation therapy (EBRT). Information on local failure was available for only 1275 patients. Clinical and treatment characteristics of patients were summarized in two categorical variables: dose (0, 1, 2, 3) and prognosis (0, 1, 2). The four dose levels are corresponding to the intervals [41.4,70.2], [71.6,75.6], [75.6,79.2] and [84.6,86.4] Gy, and three prognosis categories (low, intermediate and high risk) were defined by pre-treatment PSA, Gleason score and tumor stage.

6.1.1 Biochemical failure

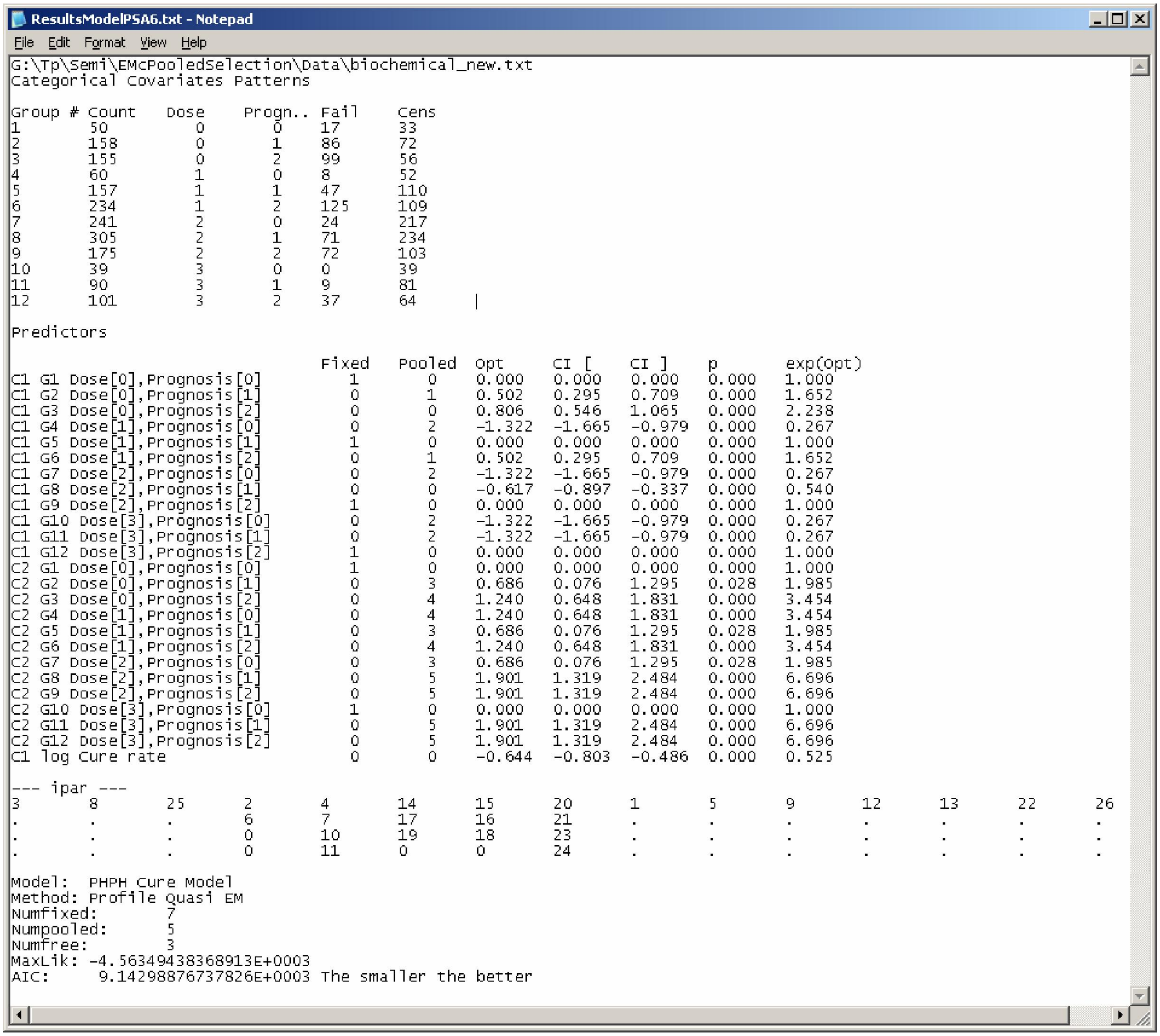
Relative risk estimates and estimated probabilities of long-term survival resulting from computer-intensive backwards pooling model selection procedure (Section 4.2.2) are shown in Table 1.

<table>
<tr><th rowspan="2">Relative Risk</th><th rowspan="2">Prognosis group</th><th colspan="4">Dose group</th></tr>
<tr><th>0</th><th>1</th><th>2</th><th>3</th></tr>
<tr><td rowspan="3">Long-Term Effect</td><td>0</td><td>1.0</td><td colspan="3">0.27</td></tr>
<tr><td>1</td><td>1.65</td><td>1.0</td><td>0.54</td><td>0.27</td></tr>
<tr><td>2</td><td>2.24</td><td>1.65</td><td colspan="2">1.0</td></tr>
<tr><td rowspan="3">Short-Term Effect</td><td>0</td><td>1.0</td><td>3.45</td><td>1.99</td><td>1.0</td></tr>
<tr><td>1</td><td colspan="2">1.99</td><td colspan="2">6.70</td></tr>
<tr><td>2</td><td colspan="2">3.45</td><td colspan="2">6.70</td></tr>
<tr><td rowspan="3">Probability of long-term survival</td><td>0</td><td>0.59</td><td colspan="3">0.87</td></tr>
<tr><td>1</td><td>0.42</td><td>0.59</td><td>0.75</td><td>0.87</td></tr>
<tr><td>2</td><td>0.31</td><td>0.42</td><td colspan="2">0.59</td></tr>
</table>

Table 1: Relative risk estimates for long- and short-term effects for the final model for biochemical failure endpoint. MSKCC database analysis.

Shown in the following Figure 16 is a detailed EMc package output that was used to build the table.

Based on these result, we can make the following key conclusions (all effects are highly significant, see Figure 16 for details).



```
G:\Tp\Semi\EMcPooledSelection\Data\biochemical_new.txt
Categorical Covariates Patterns

Group # Count   Dose    Progn.. Fail    Cens
1       50         0       0    17      33
2       158        0       1    86      72
3       155        0       2    99      56
4       60         1       0    8       52
5       157        1       1    47      110
6       234        1       2    125     109
7       241        2       0    24      217
8       305        2       1    71      234
9       175        2       2    72      103
10      39         3       0    0       39
11      90         3       1    9       81
12      101        3       2    37      64

Predictors

                                Fixed   Pooled  opt     CI [    CI ]    p       exp(opt)
C1 G1 Dose[0],Prognosis[0]         1       0    0.000   0.000   0.000   0.000   1.000
C1 G2 Dose[0],Prognosis[1]         0       1    0.502   0.295   0.709   0.000   1.652
C1 G3 Dose[0],Prognosis[2]         0       0    0.806   0.546   1.065   0.000   2.238
C1 G4 Dose[1],Prognosis[0]         0       2    -1.322  -1.665  -0.979  0.000   0.267
C1 G5 Dose[1],Prognosis[1]         1       0    0.000   0.000   0.000   0.000   1.000
C1 G6 Dose[1],Prognosis[2]         0       1    0.502   0.295   0.709   0.000   1.652
C1 G7 Dose[2],Prognosis[0]         0       2    -1.322  -1.665  -0.979  0.000   0.267
C1 G8 Dose[2],Prognosis[1]         0       0    -0.617  -0.897  -0.337  0.000   0.540
C1 G9 Dose[2],Prognosis[2]         1       0    0.000   0.000   0.000   0.000   1.000
C1 G10 Dose[3],Prognosis[0]        0       2    -1.322  -1.665  -0.979  0.000   0.267
C1 G11 Dose[3],Prognosis[1]        0       2    -1.322  -1.665  -0.979  0.000   0.267
C1 G12 Dose[3],Prognosis[2]        1       0    0.000   0.000   0.000   0.000   1.000
C2 G1 Dose[0],Prognosis[0]         1       0    0.000   0.000   0.000   0.000   1.000
C2 G2 Dose[0],Prognosis[1]         0       3    0.686   0.076   1.295   0.028   1.985
C2 G3 Dose[0],Prognosis[2]         0       4    1.240   0.648   1.831   0.000   3.454
C2 G4 Dose[1],Prognosis[0]         0       4    1.240   0.648   1.831   0.000   3.454
C2 G5 Dose[1],Prognosis[1]         0       3    0.686   0.076   1.295   0.028   1.985
C2 G6 Dose[1],Prognosis[2]         0       4    1.240   0.648   1.831   0.000   3.454
C2 G7 Dose[2],Prognosis[0]         0       3    0.686   0.076   1.295   0.028   1.985
C2 G8 Dose[2],Prognosis[1]         0       5    1.901   1.319   2.484   0.000   6.696
C2 G9 Dose[2],Prognosis[2]         0       5    1.901   1.319   2.484   0.000   6.696
C2 G10 Dose[3],Prognosis[0]        1       0    0.000   0.000   0.000   0.000   1.000
C2 G11 Dose[3],Prognosis[1]        0       5    1.901   1.319   2.484   0.000   6.696
C2 G12 Dose[3],Prognosis[2]        0       5    1.901   1.319   2.484   0.000   6.696
C1 log Cure rate                   0       0    -0.644  -0.803  -0.486  0.000   0.525

--- ipar ---
3       8       25      2       4       14      15      20      1       5       9       12      13      22      26
.       .       .       6       7       17      16      21      .       .       .       .       .       .       .
.       .       .       0       10      19      18      23      .       .       .       .       .       .       .
.       .       .       0       11      0       0       24      .       .       .       .       .       .       .

Model:  PHPH Cure Model
Method: Profile Quasi EM
Numfixed:       7
Numpooled:      5
Numfree:        3
MaxLik: -4.56349438368913E+0003
AIC:     9.14298876737826E+0003 The smaller the better
```

Figure 16: Output of the EMc package representing the final model for biochemical failure.

1. Overall, increasing dose improves the chance of long-term survival of localized disease patients with any prognosis.

2. In the favorable prognostic group dose levels 2 and 3 represent a possible over-treatment. This is evident from the fact that long-term effects for dose levels 1,2, and 3 were pooled in search for the best model.

3. Prognosis groups 1 and 2 show more rapid development of failure with dose increasing from level 1 to 2. This leads to the speculation that 75.6 Gy is a possible cutpoint when dose starts affecting the dynamics of tumor re-growth after radiation therapy.

Shown in Figure 17 is a regression tree built using the methodology developed in this project (Section 4.2.3). While the tree does not allow us to directly test statistical hypotheses, it provides a data mining tool that can suggest a direction for further clinical trials. For example as evident from Figure 17, lowest AIC corresponds to nodes 5 and 6. These nodes represent a partition of dose for low PSA high grade patients (node 5), and higher PSA early stage patients (node 6), which indicates possible subgroups of patients where increasing radiation dose may have the highest effect.

6.1.2 Local failure

Relative risk estimates and estimated probabilities of long-term survival resulting from computer-intensive backwards pooling model selection procedure (Section 4.2.2) applied to the local failure endpoint are shown in Table 2.

Relative Risk	Prognosis group	Dose group			
		0	1	2	3
Long-Term Effect	0	1.0			
	1	2.89			1.0
	2	2.89		1.0	
Short-Term Effect	0	1.0			
	1	2.85	1.0		
	2	2.85			
Probability of long-term survival	0	0.93			
	1	0.81			
	2	0.81		0.93	

Table 2: Relative risk estimates for long- and short-term effects for the final model for the local failure endpoint. MSKCC database analysis.

From the table it is clear that the local failure endpoint is much less sensitive to the treatment dose than a biochemical failure. This could be explained in part by the smaller number of events observed for this endpoint and the associated power deficiency. Based on the table we may suggest the following conclusions.

1. Overall, increasing dose is associated with a highly significant improvement in the long-term survival.

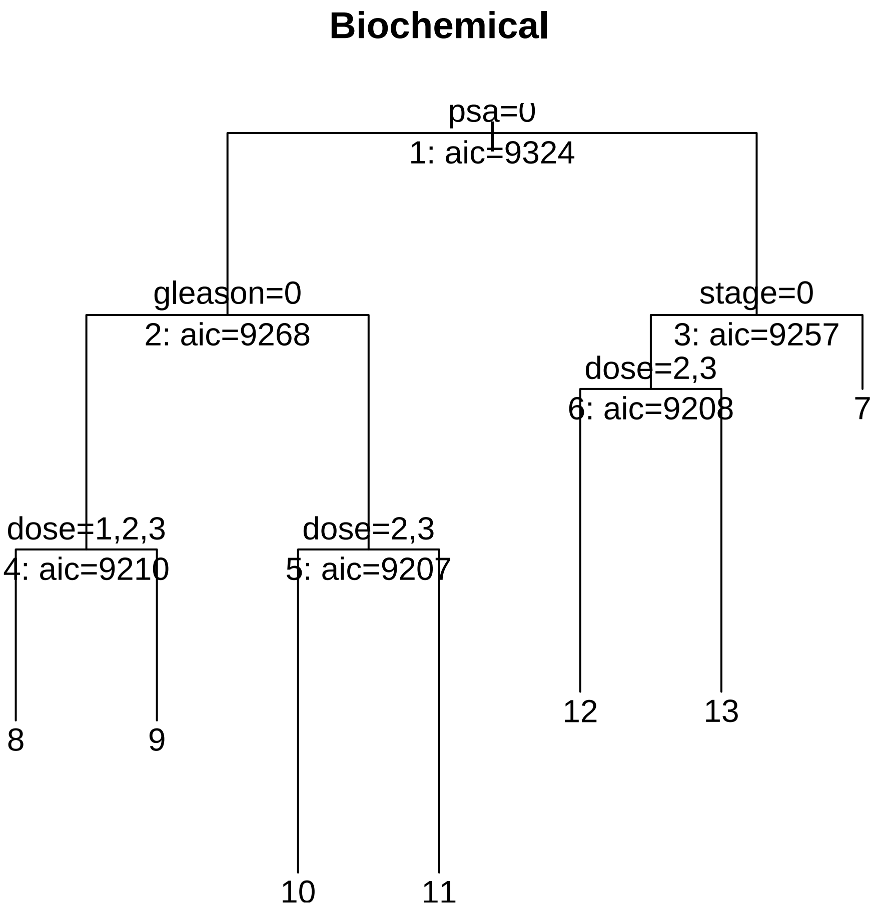


Figure 17: Regression tree built for biochemical failure endpoint

2. Dose levels 1,2, and 3 may represent an over-treatment of patients with the most favorable prognosis. This is evident from the fact that all dose levels were pooled together in long- and short-term predictors for the favorable prognosis group (prognosis=0).

3. Patients with intermediate or unfavorable prognosis benefit primarily from highest dose levels.

4. Although the short-term effect is significant, we find it difficult to interpret due to a lack of clear regularity.

Local failure endpoint is not informative enough to build a meaningful regression tree.

6.1.3 Distant metastasis

Relative risk estimates and estimated probabilities of long-term survival resulting from computer-intensive backwards pooling model selection procedure (Section 4.2.2) applied to the distant metastasis endpoint are shown in Table 3.

Relative Risk	Prognosis group	Dose group			
		0	1	2	3
Long-Term Effect	0	1.0			
	1	7.32		3.69	
	2	25.3	18.7		
Short-Term Effect	0	1.0			
	1				
	2				
Probability of long-term survival	0	0.82			
	1	0.82		0.92	
	2	0.56	0.65		

Table 3: Relative risk estimates for long- and short-term effects for the final model for the distant metastasis endpoint. MSKCC database analysis.

Based on the table we may suggest the following conclusions.

1. Overall, increasing dose is associated with a reduced chance to develop distant metastasis (highly significant).

2. No dose effect on metastasis is observed for patients with favorable prognosis. This may be due to a lack of power as such patients generally have a low chance to develop metastases.

3. There is no short-term effect, and the optimal model for metastasis endpoint is essentially a proportional hazards model.

Shown in Figure 18 is a regression tree built using the methodology developed in this project (Section 4.2.3). As is evident from the figure, there is a well structured prognostic groups subdivision for early stage patients. Also the tree suggests that increasing dose might benefit early stage high grade patients.

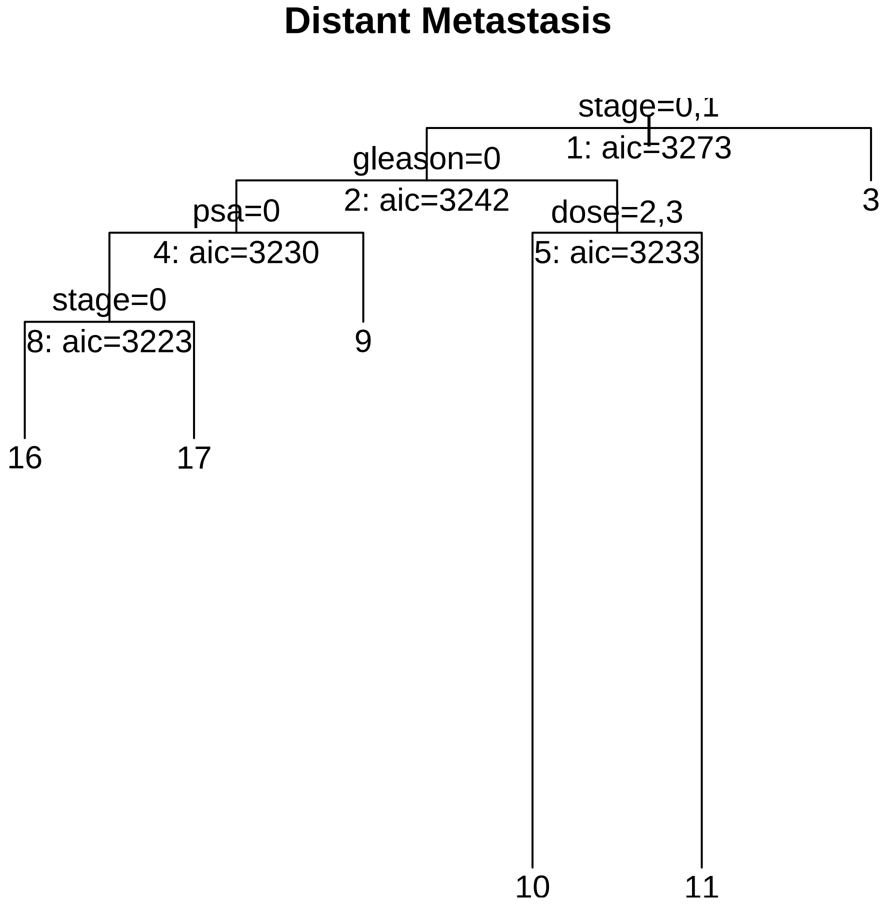


Figure 18: Regression tree built for the distant metastasis endpoint

6.1.4 Cause-specific survival

Relative risk estimates and estimated probabilities of long-term survival resulting from computer-intensive backwards pooling model selection procedure (Section 4.2.2) applied to the cause-specific survival endpoint are shown in Table 4.

Relative Risk	Prognosis group	Dose group			
		0	1	2	3
Long-Term Effect	0	1.0			
	1	1.0			
	2	1.9	4.02		
Short-Term Effect	0	1.0			
	1	9.64			
	2	28.79	9.64		
Probability of long-term survival	0	0.93			
	1	0.82			
	2	0.69	0.46		

Table 4: Relative risk estimates for long- and short-term effects for the final model for the cause-specific survival endpoint. MSKCC database analysis.

Based on the table we may suggest the following conclusions.

1. Except for patients with unfavorable prognosis, no dose effect is observed on the prostate-specific long-term survival. This may be due to lack of power.

2. A controversial adverse long-term effect of dose is observed in the unfavorable prognosis group. This finding stands in need of explanation.

3. Patients unfavorable prognosis show short-term benefit associated with increased treatment dose.

Shown in Figure 19 is a regression tree built using the methodology developed in this project (Section 4.2.3). The only node involving dose has a suboptimal AIC, which suggests that there might not be a dose effect on cause-specific survival.

6.2 SEER public database

SEER public database offers an opportunity to study the effects of surgery and radiation in subsets of patients defined by stage and grade. In order to avoid confounding due to dissemination of the PSA test in the population in the late 80ies and lead-time, length bias and overdiagnostic bias that dramatically affect survival curves during this transient period, we focused on cases diagnosed after 1990. A subset of 23,606 such cases diagnosed in the San-Francisco area was selected for the analysis. Since a combination of surgery and radiotherapy is very uncommon in prostate cancer, we do not show estimates pertaining to this group of patients.

Relative risk estimates resulting from computer-intensive backwards pooling model selection procedure (Section 4.2.2) applied to the cause-specific survival endpoint are shown in Table 5.

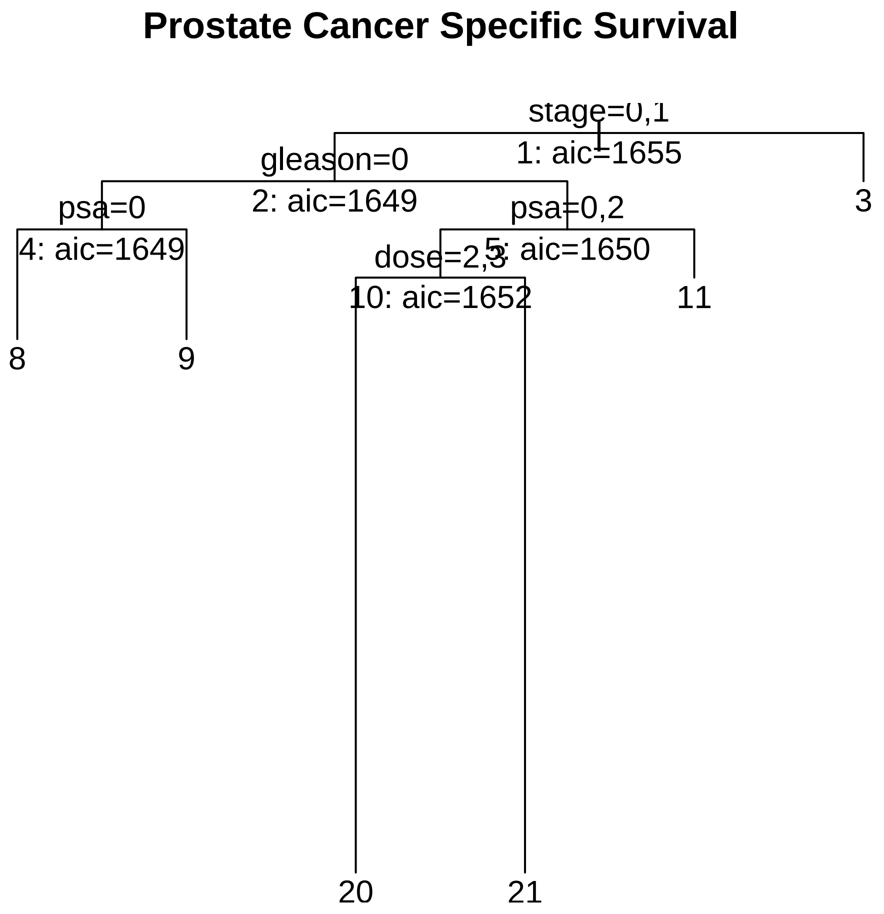


Figure 19: Regression tree built for the prostate cancer-specific survival endpoint

Stage		Localized/Regional			Distant		
Treatment		No RT No Surg	No RT Surgery	RT No Surg	No RT No Surg	No RT Surgery	RT No Surg
Long-term Effect	Low Grade	1.0	0.19	2.42	1.0	0.41	1.0
	High Grade	3.96	1.0	2.42	1.83	0.41	1.83
Short-term Effect	Low Grade	1.0	1.0	0.25	1.0	1.0	1.93
	High Grade	1.52	0.48	1.0	1.0	1.93	

Table 5: Relative risk estimates for long- and short-term effects for the final model for the cause-specific survival endpoint. SEER database analysis.

Based on the table we may suggest the following conclusions.

1. Surgery shows long-term advantage over no treatment, presumably watchful waiting, accross stages and grades.

2. A controversial adverse long-term effect of radiation is observed in low grade patients. A possible explanation might be that the effect is confounded by PSA. PSA measurements are only available in SEER for a couple of recent years. The decision to treat with radiotherapy may show a positive correlation with PSA levels at diagnosis, and therefore low grade localized stage patients treated by radiotherapy may have higher PSA levels than similar watchful waiting patients.

3. Long-term survival rates for high grade localized patients treated by surgery (predominantly radical prostatectomy) are superior to those of radiotherapy.

4. Both surgery and radiotherapy also show a short-term advantage in high-grade localized tumors.

5. In low-grade localized patients, only surgery shows a short-term advantage.

6. Surgery is the only treatment showing an effect in SEER distant stage.

7. Short-term effect of either surgery or radiation in distant stage is an adverse one.

7 Key Research Accomplishments

Simmarizing, the key research accomplishments of the project are:

1. Development of the class of Nonlinear Transformation Models (NTM) and associated QEM estimation procedures and their computer implementation;

2. Development of composition technique as a tool for model building.

3. Development of a numerically efficient algorithm for estimation of the inverse of profile information matrix for the models that paved the way to efficient model selection and hypothesis testing procedures.

4. Development of shareware software that makes these powerful tools available for broad scientific community.

5. Multivariate computer-intensive regression analysis of two large prostate cancer databases that allowed us to draw many non-trivial conclusion about the efficacy of prostate cancer treatment in various subsets of patients defined by clinical and prognostic characteristics of their tumors (Section 6)

8 Reportable Outcomes

8.1 Manuscripts

1. Tsodikov, A. (2003) Semiparametric models: A generalized self-consistency approach, *Journal of the Royal Statistical Society*, Series B, Vol. 65, 759-774.

2. Tsodikov, A., Ibrahim, J.G., and Yakovlev, A.Y. (2003) Estimating Cure Rates from Survival Data: An Alternative to Two-Component Mixture Models, *Journal of the American Statistical* Association, Vol. 98, 1063-1078.

3. Boucher, K., Asselain, B., Tsodikov, A., Yakovlev, A. (2004) Semiparametric versus parametric regression analysis based on the Bounded Cumulative Hazard Model: An application to breast cancer recurrence (invited paper), In Semiparametric Models in Survival Analysis, Quality of Life and Reliability Series: Statistics for Industry and Technology, Nikulin, M.S., Balakrishnan, N., Mesbah, M., Limnios, N. (Eds.), 2004, XLIV, 556 p. 38 illus., Hardcover, ISBN: 0-8176-3231-X, A Birkhäuser book.

4. Tsodikov, A. (2004) Generalized self-consistency methods for cure models, In "Recent developments in censored data analysis" INSERM, Paris, 2004.

5. Broët, P., Tsodikov, A., De Rycke, Y., Moreau, T. (2004) Two-sample statistics for testing the equality of survival functions against improper semi-parametric accelerated failure time alternatives: An application to the analysis of a breast cancer clinical trial, Lifetime Data Analysis, Vol. 10, 103-120.

6. Wendland, M.M., Tsodikov, A., Glenn, M.J., Gaffney, D.K. (2004) Time interval to the development of breast cancer following treatment for Hodgkin's Disease, Cancer, Vol. 101, 1275-1282.

7. Tsodikov, A., Szabo, A., and Wegelin, J. (2006) A population model of prostate cancer incidence, *Statistics in Medicine*, in press.

8. Tsodikov, A. Compound Semiparametric Survival Models, *Biometrika*, under revision.

9. Tsodikov, A. and Garibotti, G. Profile Information Matrix for Nonlinear Transformation Models, Lifetime Data Analysis, revised, under review.

8.2 Presentations

1. Tsodikov, A. (2003) Generalized Self-Consistency Methods for Cure Models, Joint Statistical Meetings, Invited session on Cure Models. (invited), San Francisco, August 2003.

2. Tsodikov, A. (2004) Cure Models (invited), Workshop of the French National Institutes of Health (INSERM).

3. Tsodikov, A. (2004) Modeling and estimation of cancer incidence and mortality under variable dissemination of screening with application to prostate cancer (invited), International Biometric Conference, Cairns, Australia July 2004.

4. Tsodikov, A. (2004) Population impact of PSA testing. The Tenth Annual Cancer Research Symposium October 20-21, UCD Cancer Center.

5. Tsodikov, A. (2004) Modeling and estimation of cancer incidence and mortality under variable dissemination of screening with application to prostate cancer (invited), International Biometric Conference, Cairns, Australia July 2004.

6. Tsodikov, A. (2004) Population impact of PSA testing. The Tenth Annual Cancer Research Symposium October 20-21, UCD Cancer Center.

7. Tsodikov, A. (2005) The use of modeling to prove outcomes: PSA really does save lives, K30 Program Retreat, Sacramento Hilton Hotel, Program on Prostate Cancer and Metabolomics

8. Tsodikov, A. (2005) invited discussion, Biomarkers in Cancer, Mathematical Biosciences Institute (MBI) at Ohio State University, Columbus, April 20-22, 2005.

9. Tsodikov, A. (2005), Invited seminar ”Computational approaches to semiparametric models”, MD Anderson Cancer Center, Houston, April 6, 2005

10. Tsodikov, A. (2005) Modeling and estimation of trends in cancer incidence and mortality with application to prostate cancer (invited), Joint Statistical Meetings, Minneapolis.

9 Conclusions

We have completed methodology and software development for point and interval estimation and variable selection for compound Nonlinear Transformation Models. We have built a number of candidate compound models for prostate cancer and verified their properties analytically and by simulations. We used the new software and methodology to apply these models to a number of real and simulated test data sets. This methodology and software arsenal was used to identify subsets of patients showing varying treatment effects. Broadly speaking, we found that increasing radiotherapy dose is benefitial for biochemical recurrence, local failure and distant metastasis, and has the potential to improve long-term survival except perhaps in a subgroup of patients with most favorable prognosis. However, no proven benefit was discovered as far as cause-specific survival is concerned, which rases the question of whether improving local control in prostate cancer is an optimal strategy to reduce mortality from the disease. Analysis of population registry data indicates that radical prostatectomy may have an advantage over radiotherapy in some subsets of patients with localized disease. However, this result may be confounded by missing PSA data in SEER registry.

10 List of Personnel

The following list represents personnel participating in this project at different stages. After the transfer of the grant from University of Utah to UC Davis, some Co-Investigators were turned into consultants to ensure continuity of the research effort. Changes in project personnel within the institutions were done in pursuit of the best expertise required at different stages of project development.

1. Alex Tsodikov, Ph.D., Professor of Biostatistics, Principal Investigator (All aspects of the project);

2. Gilda Garibotti, Ph.D., Research Associate (Utah), Consultant (Davis) (R-package implementing methodology developed in the project, profile information matrix and standard errors);

3. Andrei Yakovlev, Ph.D., Professor and Chair of Biostatistics, Co-Investigator (Utah), Consultant (Davis) (Biological interpretation of the models and mechanistic modeling, interpretation of data analyses);

4. Chuck Wiggins, Research Assistant Professor, Director, Utah Cancer Registry, Co-Investigator (Utah) (Analysis and management of cancer registry data);

5. Leonid Hanin, Professor of Mathematics, Consultant (Mechanistic models of prostate cancer, model identifiability);

6. Marco Zaider, Professor of Medical Physics, Consultant (Sloan Kettering Cancer Center database, development of models, interpretation and hypothesis-driven strategy of data analysis, expertise in Radiotherapy for the disease);

7. Ralph deVere-White, Mb, Bch, BAO., Co-investigator (Davis), Professor of Urology, Director, UC Davis Cancer Center (Clinical expertise in prostate cancer treatment and outcomes, interpretation and conclusions from data analysis, reporting and publications);

8. Jake Wegelin, Ph.D., Adjunct Assistant Professor, Co-Investigator (Davis), (standard statistical analysis, R-programming);

9. Ying-Fang Wang, Graduate Research Assistant (Davis) (Cleaning of the data, data quality control, data management and descriptive statistics)

10. Szu-Ching Tseng, Graduate Research Assistant (Software implementation of the models and methods in Delphi, data analysis)

11. Hao Liu, Adjunct Assistant Professor of Biostatistics (Davis), replacement for Jake Wegelin (asymptotic theory and properties of the semiparametric estimators)

12. Danh Nguyen, Adjunct Assistant Professor of Biostatistics (Davis), replacement for Jake Wegelin (Regression Trees methodology, R-programming)

11 References

References

L. Breiman, T.L. Friedman, R.A. Olshen, and D. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

M.M. Wendland, A. Tsodikov, M.J. Glenn, and D.K. Gaffney. Time interval to the development of breast cancer following treatment for hodgkin's disease. *Cancer*, 101:1275–1282, 2004.

12 Appendix

List of papers presented in the appendix

1. Tsodikov, A. Compound Semiparametric Survival Models, *Biometrika*, under revision.
2. Tsodikov, A. and Garibotti, G. Profile Information Matrix for Nonlinear Transformation Models, Lifetime Data Analysis, revised, under review.
3. Tsodikov, A., Szabo, A., and Wegelin, J. (2006) A population model of prostate cancer incidence, *Statistics in Medicine*, in press.

Compound Semiparametric Survival Models

A. Tsodikov

December 19, 2004

University of California, Department of Public Health Sciences,
Division of Biostatistics, One Shields Avenue,
Davis, CA 95616, U.S.A.,
atsodikov@ucdavis.edu

Abstract

Introducing a random effect into the Cox model is a useful tool for building hierarchical families of univariate semiparametric regression survival models. Hougaard [1984] used the Laplace transform to build frailty models with explicitly defined survival functions and random effects. The family derived from stable distributions was then extended [Aalen, 1992] to frailty variables following a Discrete–Continuous compound (Poisson–Gamma) structure. Still, in this form the techniques applies only to a subset of frailty models. In this paper we extend the idea of compounding first to arbitrary frailty models and then to non-frailty Nonlinear Transformation Models (NTM). EM algorithm can be used to provide inference with frailty models. Motivated by second moment properties of frailty models, Tsodikov [2003] generalized the EM algorithm into a non-frailty frame represented by the Quasi-EM algorithm (QEM). We derive a chain rule showing that QEM will fit any model constructed using the new composition technique, provided it is applicable to the submodels. Simulations, real data and a variety of models are used to illustrate the composition technique. Non-identifiability aspect of semiparametric frailty models is discussed. Many important modelling issues and links are highlighted.

1 Introduction

The model diversity in semiparametric regression analysis has been largely developed on a case by case basis. Models include the Proportional Hazards (PH) model [Cox, 1972], the Proportional Odds model (PO) [Bennett, 1983], generalized Odds-rate model [Dabrowska and Doksum, 1988], linear transformation models [Cheng et al., 1995], models motivated by frailties such as cure models [Tsodikov et al., 2003]. Despite efforts to built a universal framework of statistical inference with semiparametric models, including general instruments for model building, numerical estimation algorithms, identifiability and asymptotics, general and practical results are still a challenge, and most existing inferential tools are model-specific. In this paper we develop a fragment of a general approach for a class of semiparametric models equipped with model-building and inferential algorithms.

Motivated by second-order properties of frailty models Tsodikov [2003] proposed a family of so-called Nonlinear Transformation Models (NTM) and supplied it with a general numerical inference framework based on the QEM algorithm, a subset of recently developed MM algorithms [Lange et al., 2000]. In this paper we will equip the NTM-QEM frame with a model-building tool that allows us to generate a span of hierarchical models from a set of basis models such as PH and PO, such that they remain within the frame and QEM is guaranteed to work on any descendant model. When none of the basis models is suitable for the data, the technique offers an automatic way of combining features of simpler models in search of a more complex model that would be right for the data and such that inference procedures are still available.

Let $\gamma(x\,|\boldsymbol{\theta})$ be a parametrically specified strictly increasing distribution function on $[0, 1]$, where $\boldsymbol{\theta}$ is a vector of parameters. To define an NTM, we first make γ a regression model by turning $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ into a set of predictors depending on covariates, $\boldsymbol{z}$, and regression coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_k)$. Typically, $\theta_i(\boldsymbol{\beta}_i, \boldsymbol{z}) = \exp\{\boldsymbol{\beta}_i^{\mathrm{T}}\boldsymbol{z}\}$. Thus, γ explicitly represents the parametric part of the model and is called an NTM generating function. Denote by $\mathcal{N}$ the class of all NTM generating functions.

Let $F(t)$ be a nonparametrically specified (a step-function) baseline survival function. In a Nonlinear Transformation Model it is assumed that survival function $G(t\,|\,\boldsymbol{\beta}, \boldsymbol{z})$ can be represented as

$$G(t\,|\,\boldsymbol{\beta}, \boldsymbol{z}) = \gamma\left\{F(t)\,|\,\boldsymbol{\beta}, \boldsymbol{z}\right\} = (\gamma \circ F)(t\,|\,\boldsymbol{\beta}, \boldsymbol{z}). \tag{1}$$

Consider a sample of right censored data under non-informative censoring. The plug-in form (1) induces an Von-Mises style likelihood whose Fréchet derivative with respect to F leads to a self-consistency score equation

$$h_t = \frac{D_t}{\sum_{i \in \mathcal{R}_t} \Theta(F(t_i) \,|\, \boldsymbol{\beta}, \boldsymbol{z}_i, c_i)}, \tag{2}$$

where $h_t = H(t) - H(t-0)$ is the jump of the baseline cumulative hazard $H = -\log F$ at time t, $\mathcal{R}_t$ is a set of subjects at risk at time t, t_i are their event times, c is a censoring indicator ($c = 1$ for a failure, $c = 0$ for a censored observation), and Θ is a parametric function defined through γ as

$$\Theta\left[x \,|\, \cdot, c\right] = c + x \frac{\gamma^{(c+1)}(x \,|\, \cdot)}{\gamma^{(c)}(x \,|\, \cdot)}, \tag{3}$$

where $\gamma^{(i)}(x \,|\cdot)$ is the derivative of γ of the ith order with respect to x, $\gamma^{(0)} \equiv \gamma$. For a frailty model, Θ is the conditional expectation of the frailty variable given observed data and represents the result of the E-step of an EM algorithm [Tsodikov, 2003].

Solving (2) by iterations

$$H^{(k+1)} = \varphi(H^{(k)}), \tag{4}$$

where φ denotes the right part of the self-consistency equation (2) as a functional of $H = -\log F$, and k counts iterations, is an QEM algorithm. Its point of convergence represents a fixed point of φ and an NPMLE of F (or H) given $\boldsymbol{\beta}$. Note that in this form the algorithm is not based on missing data and is applicable to non-frailty models. For a frailty model though, this is an EM algorithm. It can be shown that if $\Theta(x|\cdot)$ is a non-decreasing function of x, which is the case for all frailty models, then each QEM iteration improves the likelihood, which is a key property for convergence. The solution of the self-consistency equation can be written as an implicit function of $\boldsymbol{\beta}$, $F = F(\boldsymbol{\beta})$. Plugging this solution into the full loglikelihood $\ell(\boldsymbol{\beta}, F)$ gives us the profile likelihood

$$\ell_{pr}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}, F(\boldsymbol{\beta})), \tag{5}$$

that is used to provide inference for $\boldsymbol{\beta}$, [Murphy and Van der Vaart, 2000, Tsodikov, 2003].

Aside from a broader class of models, the advantage of the NTM-QEM approach over the traditional frailty-EM one is that analytic work required to specify the algorithm and verify its applicability is minimized. Given a new model $G = \gamma(F)$, with the traditional frailty framework, one would first have to verify that γ is a completely monotonic function (Bernstein theorem, [Feller, 1971]) to ensure that the new model is a frailty model. Then one needs to invert the Laplace transform to find the distribution of the frailty random variable U. With the conditional distribution of U, given observed data, the conditional expectation of U is derived to provide missing data imputation. Closed form expressions are required throughout to ensure a numerically efficient algorithm. NTM-QEM approach makes the above exercise obsolete and boils down to taking two derivatives of γ and verifying that Θ (3) is nondecreasing. QEM is faster than the traditional EM that uses partial likelihood at the M-step even with models that have a single predictor [Tsodikov, 2003] and found applications in computer–intensive inference procedures such as the bootstrap [Dixon et al., 2005]. With models having multiple predictors or parameters, such as the Γ-frailty model, use of partial likelihood implies that parameters outside the partial likelihood still need to be estimated after EM converges, which makes the traditional frailty-EM approach very inefficient numerically.

The idea of this paper is to build semiparametric models by operation of composition of γs. Indeed, since an NTM generating function γ has the domain $[0, 1]$ and range in the same interval, a composition of any number of such functions is again an NTM generating function. Let $\gamma_i(x \mid \theta_i)$, $i = 1, 2, \ldots$, be NTM generating functions for a set of basis models. In this paper we study models built as

$$\gamma_i(x\,|\theta_i) \circ \gamma_j(x\,|\theta_j) = \gamma_i\left(\gamma_j(x\,|\theta_j)\,|\theta_i\right) = \gamma_{ij}(x\,|\theta_i, \theta_j), \tag{6}$$

from any two sumbodels γ_i and γ_j and show that QEM is applicable to any compound model. The composition techniques will be motivated by frailty models. We will show that it generalizes Aalen's compound Poisson device based on Discrete-Continuous compounding [Aalen, 1992, Moger et al., 2004] and give examples of compound frailty models. We extend the device to arbitrary (discrete or continuous) frailty submodels. We then consider composition for NTMs and discuss identifiability issues. Finally, we apply the composition technique to real data and study asymptotic properties of the profile likelihood MLEs by simulations.

2 Discrete-Continuous composition device

This section is a brief review of existing model-building methodology. We will use it in the semiparametric setting to construct the PHPH model in section 6.2.1.

Heterogeneity has been a popular tool of extending survival models. Several authors considered variations of the so-called Proportional Hazards (PH) frailty model [Hougaard, 1984, Aalen, 1992, Klein, 1992, Nielsen et al., 1992],

$$G(t) = \mathrm{E}\left\{F(t)^U\right\}, \tag{7}$$

where G is a population survival function, F is the baseline survival function, and U is a nonnegative random variable (frailty). Hougaard [1984] observed that (7) can be written as a Laplace transform of U

$$G(t) = \mathcal{L}_U\left\{H(t)\right\}, \;\; \mathcal{L}_U(s) = \mathrm{E}\left\{e^{-sU}\right\}, \tag{8}$$

where $H = -\log(F)$ is the baseline cumulative hazard.

The Laplace transform connection (8) was used by Aalen [1992], Moger et al. [2004] to build a family of models induced by a compound Poisson-Gamma distribution for U. The family was used in the parametric setting with $H(t)$ specified according to Weibull distribution.

Tsodikov et al. [2003] used the idea as an instrument to build semiparametric survival models induced by compound Discrete-Continuous frailties. The frailty random variable U was regressed on covariates $\boldsymbol{z}$, so that $U(\boldsymbol{\beta}, \boldsymbol{z})$ is considered as a response variable in a parametric regression model, where $\boldsymbol{\beta}$ is a vector of regression coefficients. The parameters of the distribution of U were thought of as regression predictors depending on covariates and regression coefficients. Let $\nu(\boldsymbol{\beta}_\theta, \boldsymbol{z})$ be a discrete nonnegative random variable with the distribution with parameter vector $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\beta}_\theta, \boldsymbol{z})$ and the Laplace transform $\mathcal{L}_{\boldsymbol{\theta}}$. Let ξ_k be i.i.d. copies of a random variable $\xi(\boldsymbol{\beta}_\eta, \boldsymbol{z})$ parameterized through $\boldsymbol{\eta}(\boldsymbol{\beta}_\eta, \boldsymbol{z})$ with the Laplace transform $\mathcal{L}_{\boldsymbol{\eta}}$. Then the compound distribution

$$U(\boldsymbol{\beta}, \boldsymbol{z}) = \sum_{k=1}^{\nu(\boldsymbol{\beta}_\theta, \boldsymbol{z})} \xi_k(\boldsymbol{\beta}_\eta, \boldsymbol{z}), \;\; \sum_1^0 = 0, \tag{9}$$

has the Laplace transform

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\eta}} = \mathcal{L}_{\boldsymbol{\theta}}\left\{-\log \mathcal{L}_{\boldsymbol{\eta}}\right\}. \tag{10}$$

In view of (8) and (10), the compound survival function generated by (9) has the form

$$G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}) = \mathcal{L}_{\boldsymbol{\theta}} \left\{ -\log \mathcal{L}_{\boldsymbol{\eta}}(H(t)) \right\}, \tag{11}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_\theta, \boldsymbol{\beta}_\eta)$. In the semiparametric context, H is treated as an infinite dimensional parameter, a step-function. As a result of the above procedure a new hierarchical model $G = \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\eta}}(H)$ is constructed so that it combines features of two submodels $G = \mathcal{L}_{\boldsymbol{\theta}}(H)$ and $G = \mathcal{L}_{\boldsymbol{\eta}}(H)$.

3 PH mixture model vs. NTM

Following [Wassel and Moeschberger, 1993, Clayton and Cuzick, 1985a] who considered frailty variables dependent on covariates, we may write a general univariate PH mixture model as

$$G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}) = \mathrm{E} \left\{ F(t)^{U(\boldsymbol{\beta}, \boldsymbol{z})} \,\middle|\, \boldsymbol{z} \right\}. \tag{12}$$

This model can be considered a generalization of the so-called PH frailty model, or a PH model with a random effect

$$G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}) = \mathrm{E} \left\{ F(t)^{\theta(\boldsymbol{\beta}, \boldsymbol{z}) V} \right\}, \tag{13}$$

where θ is a predictor, and V is a random variable independent of the covariates, considered by [Hougaard, 1984, Klein, 1992, Nielsen et al., 1992] and many other authors, for different distributions of V.

We can make the following important observations about the class of PH mixture models (12). The survival function (12) is built by composition

$$G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}) = (\gamma \circ F)(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}), \tag{14}$$

where $\gamma(x \,|\, \boldsymbol{\beta}, \boldsymbol{z})$ belongs to the class $\mathcal{P}$ of probability generating functions (p.g.f.). Here we extend the use of p.g.f. to arbitrary nonnegative random variables and define any such p.g.f. p as $p(x) = \mathcal{L}\left\{-\log(x)\right\}$, where $\mathcal{L}$ is the Laplace transform. Covariates enter $\gamma \in \mathcal{P}$ through the parameters $\boldsymbol{\theta}$ of the distribution of U as they are turned into regression predictors, typically as $\boldsymbol{\theta} = (\exp\{\boldsymbol{\beta}_1^{\mathrm{T}} \boldsymbol{z}\}, \ldots, \exp\{\boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{z}\})^{\mathrm{T}}$, $k = \dim(\boldsymbol{\theta})$. Note that p.g.f. thus defined is an NTM generating function (see Introduction), so that $\mathcal{P} \subset \mathcal{N}$, and PH mixture models (12) are a subclass of NTMs (1). Since p.g.f. is an analytic function, the complement of the class of frailty models to the NTM class includes, for example, models with a non-existent derivative $\gamma^{(k)}$ for some $k > 2$.

4 Imputation operator and the meaning of nondecreasing Θ

Suppose, we have an observation $(t, \boldsymbol{z}, c)$ sampled from the PH mixture model under independent censoring, where t is an observed survival time and c is a censoring indicator ($c = 0$ if t is a censored survival time, and $c = 1$ if t is a failure). Then, under the PH mixture model (12), the conditional expectation of U, given the observed event $(t, \boldsymbol{z}, c)$ is given by [Tsodikov, 2003]

$$\mathrm{E}\left\{U(\cdot) \,|\, t, \cdot, c\right\} = (\Theta \circ F)(t \,|\, \cdot, c) = \Theta\left[F(t) \,|\, \cdot, c\right],$$

where the function Θ is given by (3). For brevity, we use $(\cdot)$ to suppress covariates and regression coeffitients $\boldsymbol{\beta}, \boldsymbol{z}$. While Θ in the Introduction is defined for NTMs, we also consider the p.g.f. subclass $\gamma \in \mathcal{P} \subset \mathcal{N}$ as a motivation and to better understand the conditions that make the NTM-QEM tandem work.

Cauchy-Schwartz inequality can be used to show that for any mixture model $\gamma \in \mathcal{P}$, $\Theta\left[x \,|\, \cdot, c\right]$ is nondecreasing in x for any $c = 0, 1$. The nondecreasing character of the function Θ in the above statement is quite natural. The longer the subject stays event–free, the lower the subject's posterior risk, represented by Θ. So $\Theta\{F(t) \,|\, \cdot, c\}$ must be a nonincreasing function of t for both failure ($c = 1$) and censoring ($c = 0$) events. Since the survival function $F(t)$ is nonincreasing in t, $\Theta(x \,|\, \cdot, c)$ must be nondecreasing in x. It is interesting to note that the population hazard function for a heterogeneous population under the PH mixture model is expressed as $\lambda(t \,|\, \boldsymbol{z}) = \Theta\{F(t) \,|\, \cdot, 0\}h(t)$, where h is the hazard function corresponding to F. Even if $h(t)$ is increasing, the observed population hazard function may be a decreasing one through the decreasing behavior of $\Theta\{F(t)|\cdot, 0\}$ with time. This observation represents a selection effect of the risk set becoming "healthier" with time, as frail individuals leave the population. This effect was discovered and extensively studied in demography [Vaupel et al., 1979] in the context of misinterpretation of mortality trends.

With γ representing a PH mixture model $\gamma \in \mathcal{P}$, kth moments of the mixing variable U, $k = 1, 2, \ldots$, can be obtained through derivatives $\gamma^{(k)}$. Both Θ and QEM are defined using the derivatives up to second order of γ, $k = 1, 2$. Based on the above observations, NTM-QEM tandem is defined to follow second-order properties of the Frailty-EM frame. This is all that is needed to ensure the EM-like behavior of the QEM, and existence of all derivatives of γ (still a weaker assumption than that of a frailty model) is excessive for purposes of statistical inference.

As discussed in [Tsodikov, 2003], the property of non-decreasing Θ represents a generalized form of Jensen inequality on the primitive class of functions necessary to handle the QEM algorithm.

In addition to being a non-increasing function of time, the posterior risk $\mathrm{E}\{U(\cdot)\,|\,t,\cdot,c\}$ for PH mixture models ($\gamma \in \mathcal{P}$) has the following two natural properties.

1. Other things equal, the posterior risk of a failure is at least as high as a posterior risk of a censored subject $\mathrm{E}\{U(\cdot)\,|\,t,\cdot,1\} \geq \mathrm{E}\{U(\cdot)\,|\,t,\cdot,0\}$. This statement is valid in the general NTM form (see proposition below).

2. Since a censored observation at time $t = 0$ does not contribute any information on the risk, posterior risk for $t = 0$, $c = 0$ is the same as prior risk $\mathrm{E}\{U(\cdot)\}$. Expressing the mean of U through its p.g.f. $\gamma \in \mathcal{P}$, we have $\mathrm{E}\{U(\cdot)\,|\,0,\cdot,0\} = \mathrm{E}\{U\} = \gamma'(1\,|\cdot)$.

Proposition 4.1 *Surrogate of posterior risk for NTM.*
Let $\Theta(x\,|\cdot,c)$, be the function defined by (3) and induced by some NTM generating function γ, given an event $(t,\cdot,c)$ observed on a subject. Then

(A) If $\Theta(x\,|\cdot)$ is a non-decreasing function of x, then

$$\Theta(F(t)\,|\cdot,1) \geq \Theta(F(t)\,|\cdot,0) > 0 \tag{15}$$

(B) If $\gamma \in \mathcal{P}$ is a p.g.f. of some nonnegative random variable U, then

$$E\{U\,|\,t,\cdot,1)\} \geq E\{U\,|\,t,\cdot,0)\} > 0 \tag{16}$$

$$E\{U\,|\,0,\cdot,0)\} = E\{U\,|\,\cdot\} = \gamma'(1\,|\,\cdot) \tag{17}$$

The proof is given in the Appendix A.1. The graph of typical behavior of the posterior risk is given in Figure 1 based on the real data example considered in Section 8.1.

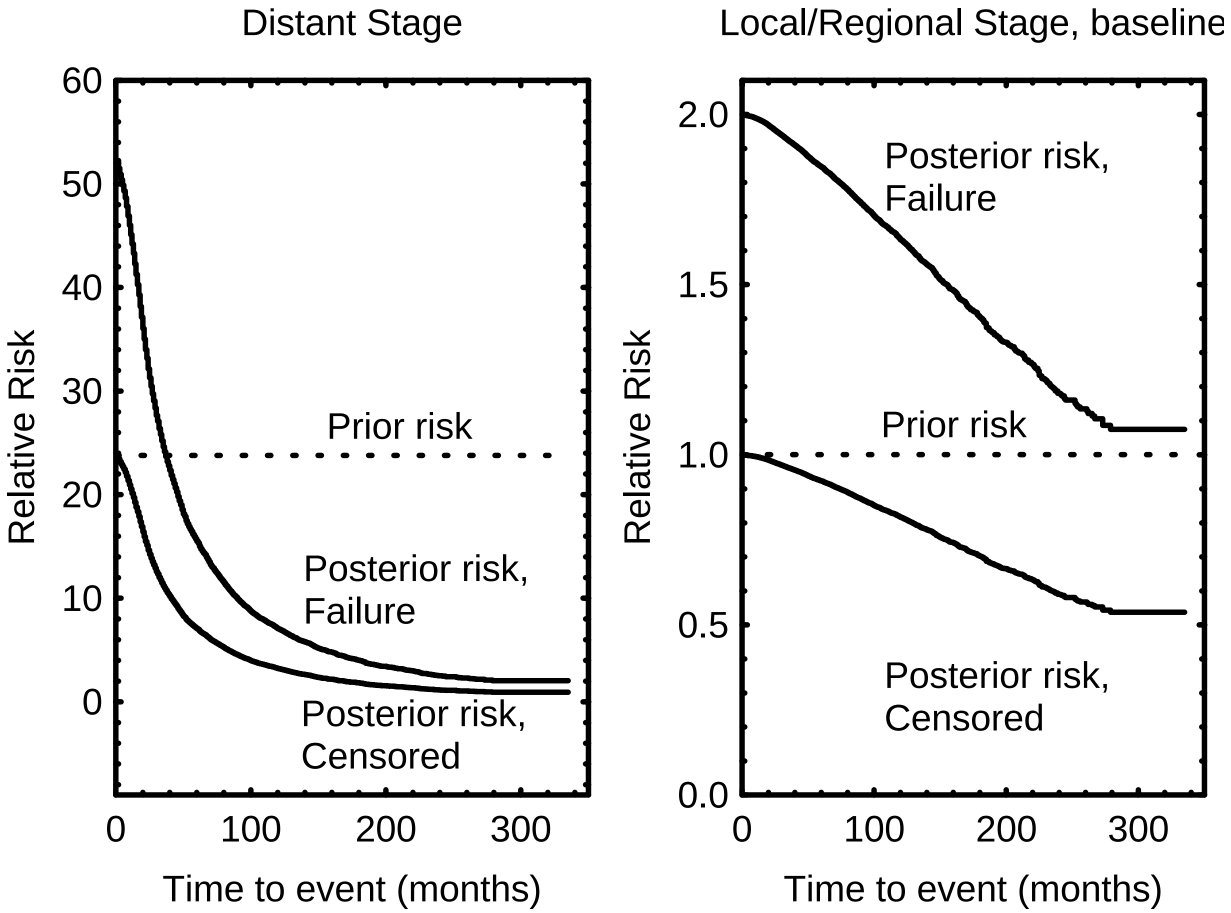


Figure 1: Posterior risk $\Theta(F(t) \,|\, \boldsymbol{\beta}, z, c)$ as a function of time to event t by type of event (failure, $c = 0$ and censoring $c = 1$), and Stage (z) (Local/Regional and Distant)

5 Model building

5.1 Composition with PH mixture models

The Discrete-Continuous composition device described in Section 2 is quite restrictive. It covers only a specific subclass of frailty models. For example, should we try to build a hierarchical model from two submodels, each generated by a continuous frailty variable, the compound device (11) would fail. With composition defined on the level of missing data (9) (random variables used to construct U), it is not clear how compositions other than of a Discrete-Continuous type could be defined. In this section we extend the device of Section 2 to arbitrary frailty models by defining the composition on the level of transforms.

Returning to Discrete-Continuous composition described by (9), observe that in terms of p.g.f. (10) can be written as

$$\gamma_{\theta,\eta} = \gamma_\theta \circ \gamma_\eta, \tag{18}$$

where γ_θ corresponds to a discrete random variable.

Now let γ_θ and γ_η be two p.g.f. of some *arbitrary* nonnegative random variables, parameterized through predictors $\theta(\boldsymbol{\beta}_\theta, \boldsymbol{z})$ and $\eta(\boldsymbol{\beta}_\eta, \boldsymbol{z})$, respectively. These functions generate two PH mixture models of the form $G(t \,|\, \boldsymbol{\beta}_\cdot, \boldsymbol{z}) = \gamma_\cdot(F(t) \,|\, \boldsymbol{\beta}_\cdot, \boldsymbol{z})$. The following proposition shows that the compound expression $\gamma_{\theta,\eta}(F \,|\, \cdot)$ is again a PH mixture model.

Proposition 5.1 *Composition for mixture models.*
Let γ_θ and γ_η be some two p.g.f. $\gamma_\theta(x|\cdot) = E(x^\nu \,|\, \cdot)$, $\gamma_\eta(x|\cdot) = E(x^\xi \,|\, \cdot)$, where ν and ξ are some independent nonnegative random variables. Then the compound function $\gamma_{\theta,\eta} = \gamma_\theta \circ \gamma_\eta$ is also a p.g.f., meaning that there exists a nonnegative random variable U such that $\gamma_{\theta,\eta}(x|\cdot) = E(x^U \,|\, \cdot)$.

The proof is given in the Appendix A.2.

The above result shows that the PH mixture subclass of NTM is closed with respect to composition of NTM generating functions, and consequently, the self-consistency equation (2) defines an EM algorithm serving the new compound model. Therefore, convergence of (4) for the compound model follows from the general EM theory [Dempster et al., 1977, Wu, 1983].

5.2 NTM composition device

In this section, we extend the composition techniques for the PH mixture model (Section 5.1) to the NTM class.

Now, let $\gamma_\theta, \gamma_\eta \in \mathcal{N}$ be two NTM generating functions with predictors θ, and η, respectively, not necessarily from the p.g.f. subclass. Then

$$\gamma(x|\cdot) = (\gamma_\theta \circ \gamma_\eta)(x|\cdot) \tag{19}$$

is a new NTM semiparametric model with two predictors θ and η. If $\gamma_\theta(x|\cdot) \equiv x$ for some value of θ (usually for $\theta = 1$), then the model (19) includes models γ_θ and γ_η as nested special cases. The following statement shows that the NTM class with non-decreasing $Theta$ is closed with respect to composition.

Proposition 5.2 *Composition chain rule for NTM.*
Let $\gamma_\theta \in \mathcal{N}$ and $\gamma_\eta \in \mathcal{N}$ be some two NTM–generating functions, each satisfying the assumption of nondecreasing Θ, where Θ is given by (3), and let $\gamma_{\theta,\eta} = \gamma_\theta \circ \gamma_\eta$ be the compound function. Let Θ_a be the Θ–function (3) corresponding to γ_a, for any a. Then
(A)

$$\Theta_{\theta,\eta}(x \mid \cdot, c) = \Theta_\eta(x \mid \cdot, 0) \left\{ \left(\Theta_\theta \circ \gamma_\eta\right)(x \mid \cdot, c) - c \right\} + c\Theta_\eta(x \mid \cdot, c), \tag{20}$$

where $c = 0, 1$ and $(\Theta \circ \gamma)(x \mid \cdot, c)$ is understood as $\Theta\{\gamma(x \mid \cdot)|\cdot, c\}$; and
(B) The compound function $\Theta_{\theta,\eta}(x \mid \cdot)$ derived from the compound NTM–generating function $\gamma_{\theta,\eta}$ is nondecreasing in x.

Proof is given in the Appendix A.3.

Equation (20) represents a chain rule for Θ for compound models and simplifies derivation of compound Θ through direct use of Θs corresponding to submodels participating in the composition.

Also, operation of composition (19) preserves the property of nondecreasing Θ discussed in Section 4. Therefore, convergence of the QEM algorithm (4) for the compound model follows from the results presented in [Tsodikov, 2003].

6 Examples of models

6.1 Basic submodels

In this section we present some popular models with one predictor that will be used as a basis to generate compound models in the next section.

6.1.1 PO model

Semiparametric PO model G (a survival function) can be defined as the one with log odds ratios of survival independent of time and an arbitrary baseline survival function F. It has long been known that the proportional odds (PO) model has frailty interpretation [Clayton and Cuzick, 1985b]. As we will see later, the above definition identifies a family of frailty models with infinitely many possible frailty distributions.

First, consider interpretation of the PO model as a geometric frailty model. Let F be the baseline survival function corresponding to the covariate vector $\boldsymbol{z}_0$ such that $\theta(\boldsymbol{\beta}, \boldsymbol{z}_0) = 1$. For the PO model, the predictor $\theta(\boldsymbol{\beta}, \boldsymbol{z})$ has the meaning of odds ratio of an observation with covariate vector $\boldsymbol{z}$, relative to the baseline. The PO assumption

$$\frac{\mathrm{Odds}\{G(t|\boldsymbol{\beta}, \boldsymbol{z})\}}{\mathrm{Odds}\{F(t)\}} = \theta(\boldsymbol{\beta}, \boldsymbol{z}), \tag{21}$$

where $\mathrm{Odds}(a) = a/(1-a)$, yields the PO model of the form $G = \gamma \circ F$, where

$$\gamma(x|\cdot) = \frac{\theta(\cdot)x}{1 - \bar{\theta}(\cdot)x}, \tag{22}$$

and $\bar{a} = 1 - a$ for any a. To invert the transform (22), we expand it in a Taylor power series about $x = 0$. We have

$$\gamma(x|\cdot) = \sum_{k=1}^{\infty} \theta(\cdot)\bar{\theta}^{k-1}(\cdot)x^k.$$

We note that the coefficients of the power series represent geometric probabilities, given $\theta(\cdot) \leq 1$, and therefore $\gamma(x|\cdot) = \mathrm{E}(x^U)$, where U is geometrically distributed.

Now, let us derive the interpretation of the PO model as an exponential frailty model. Consider a PO model of the form

$$G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}) = \frac{\theta(\boldsymbol{\beta}, \boldsymbol{z})}{\theta(\boldsymbol{\beta}, \boldsymbol{z}) + H(t)}, \quad \gamma(x \,|\cdot) = \frac{\theta(\cdot)}{\theta(\cdot) - \log x} \tag{23}$$

where H is some nonparametrically specified baseline cumulative hazard. As with the PO model (22), for any two values of the predictor, θ_1, θ_2, and the corresponding survival functions $G_i(t) = G(t \,|\, \theta_i)$, $i = 1, 2$, the odds ratio derived from (23)

$$\frac{\mathrm{Odds}\{G_1(t)\}}{\mathrm{Odds}\{G_2(t)\}} = \frac{\theta_1}{\theta_2}$$

is a constant in t, and the PO assumption is satisfied. For the PO model in the form (23), we have

$$\mathcal{L}_\theta(H) = \gamma(e^{-H}|\cdot) = \frac{\theta(\cdot)}{\theta(\cdot) + H}.$$

This is the Laplace transform of an exponential distribution with parameter $\theta(\cdot)$. Therefore in this case U follows an exponential regression model with $\mathrm{E}U = \theta(\boldsymbol{\beta}, \boldsymbol{z})^{-1}$.

While it is generally believed that frailty distributions are identifiable [Ebbers and Ridder, 1982], the observation that the two different frailty distributions lead to the same PO model illustrates an important point of non-identifiability of frailty distributions in *semiparametric* frailty models. In fact, $\boldsymbol{\beta}$s representing the log-odds ratios in both models (23) and (22) are exactly the same. This non-identifiability is due to the fact that a semiparametric model is defined as a *class* when we say that the baseline survival function is arbitrary. This class can be represented in an infinite number of equivalent forms. The models (23) and (22) use two different forms of an arbitrary baseline survival function. Indeed, if $F(t)$ is an arbitrary survival function, so is $(1 - \log F(t))^{-1}$. Obviously the choice of this form along with the parameterization of γ determines the distribution of the frailty variable as the inverse transform. We will return to this discussion in Section 7.

In order to specify the estimation algorithm, we need to derive the posterior risk function for the model. Using (3) and (23), we get

$$\Theta(x \,|\, \cdot, c) = \frac{c + 1}{\theta(\cdot) - \log x}. \tag{24}$$

6.1.2 PH models

The traditional PH model, later referred to as the Proper PH model,

$$G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}) = F(t)^{\theta(\boldsymbol{\beta}, \boldsymbol{z})} \tag{25}$$

does not need an introduction. Direct use of (3) leads us to the posterior risk function of the form

$$\Theta_\theta(x \,|\, \cdot, c) \equiv \theta(\cdot). \tag{26}$$

Note that since Θ does not depend on the infinite dimensional part of the model F, the QEM procedure and the self-consistency equation (2) reduce to the Nelson-Aalen-Breslow estimator for the baseline hazard in the PH model.

Along with (25), we consider a PH model with cure (Improper PH model). While it is sometimes believed that the chance of cure defies proportional hazards and implies a mixture model of the form $G = p(\beta, \boldsymbol{z}) + \bar{p}(\beta, \boldsymbol{z})F$, where p is the probability of cure [Kuk and Chen, 1992], this is not the case. The PH model with cure was motivated by the intention to combine proportional hazards and an improper survival function [Tsodikov, 1998]. An improper survival function $G(t)$ implies that the cumulative hazard has an asymptote, θ, as $t \to \infty$. Any such hazard can be represented as $\theta(1 - F(t))$, where $F(t)$ is a proper survival function. Introducing covariates into the parameter $\theta = \theta(\boldsymbol{\beta}, \boldsymbol{z})$ leads to a PH model with an asymptote

$$G(t \,|\, \boldsymbol{\beta}, \boldsymbol{z}) = \exp\left\{-\theta(\boldsymbol{\beta}, \boldsymbol{z})[1 - F(t)]\right\}. \tag{27}$$

Expanding the NTM generating function of the Improper PH model

$$\gamma(x) = \exp\{\theta(1 - x)\} \tag{28}$$

in a Taylor power series about $x = 0$, we obtain a power series with Poisson probabilities with parameter θ. Therefore, (27) is a Poisson frailty model. Again, we note a non-identifiability issue as non-random frailty and Poisson frailty lead to the PH model with the same hazard ratios.

Using (3) we get the posterior risk function

$$\Theta_\theta(x \,|\, \cdot, c) = c + \theta(\cdot)x. \tag{29}$$

6.1.3 Linear Transformation Models

The PO and the PH model considered above are members of the so-called linear transformation model (LTM) family defined as [Cheng et al., 1995, 1997]

$$\log v(T|\boldsymbol{z}) = -\log\theta(\boldsymbol{z}) + \epsilon, \tag{30}$$

where T is the failure time, ϵ is the random error with the distribution μ, and v is some unspecified strictly increasing function. For the exponential predictor $\theta(\boldsymbol{\beta}, \boldsymbol{z}) = \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})$, the model assumes a linear form in covariates and transformed response. The connection between LTM, the PH model, the PO model, and binary regression models was discussed in [Doksum and

Gasko, 1990]. After a little algebra, a linear transformation model can be represented as an NTM with the NTM–generating function

$$\gamma(x \mid \boldsymbol{\beta}, \boldsymbol{z}) = p \left\{ \log \theta(\boldsymbol{\beta}, \boldsymbol{z}) + q(x) \right\}, \tag{31}$$

where p is a parametrically specified tail function $(-\infty, \infty) \to [0, 1]$, (=1-distribution function), and q is an inverse tail function. It is convenient to specify q as the inverse of p, then $\theta = 1$ corresponds to the baseline $\gamma(x|\cdot) = x$.

6.2 Compound models

While the models considered below were introduced on a case by case basis and motivated by various non-systematic considerations, in this paper we re-invent them using composition as a tool to illustrate the method.

6.2.1 PHPH Cure Model

This model extends the Improper PH model (27) by introducing a PH short-term effect on the normalized baseline cumulative hazard $F \to F^{\eta(\boldsymbol{\beta},\boldsymbol{z})}$,

$$G(t \mid \boldsymbol{\beta}, \boldsymbol{z}) = \exp\left\{ -\theta(\boldsymbol{\beta}, \boldsymbol{z})[1 - F(t)^{\eta(\boldsymbol{\beta},\boldsymbol{z})}] \right\}. \tag{32}$$

Here we note that the model is constructed by composition (19) of NTM generating functions for the Improper PH model (28) and the Proper PH model $\gamma_\eta(x) = x^\eta$,

$$\begin{aligned} \gamma_{\theta,\eta}(x \mid \cdot) = \gamma_\theta(x \mid \cdot) \circ \gamma_\eta(x \mid \cdot) = \\ [\exp\{-\theta(\cdot)(1 - x)\}] \circ \left[x^{\eta(\cdot)}\right] = \exp\left\{-\theta(\cdot)\left(1 - x^{\eta(\cdot)}\right)\right\}. \end{aligned} \tag{33}$$

A review and history of this model is presented in [Tsodikov et al., 2003].

Note that based on Section 6.1.2, γ_θ is a p.g.f. of a Poisson random variable, and γ_η is a p.g.f. of a nonrandom variable. Therefore the composition is a particular case of Aalen's device [Aalen, 1992] (9) with ν being Poisson(θ), and $\xi = \eta$ being nonrandom.

Rather than compute the conditional expectation of the frailty variable using an integral over the compound distribution, we can use the chain rule (20) with the submodel-Θs specified by (26) and (29) and immediately get

$$\Theta(x|\cdot, c) = \theta(\cdot)\eta(\cdot)x^{\eta(\cdot)} + c\eta(\cdot). \tag{34}$$

6.2.2 Γ-frailty model

Now, consider a model composed of the PH and the PO models.

The Γ–frailty model can be built as a composition of the NTM–generating functions corresponding to the PO (**??**) and the proper PH models (25). As a result of the composition $\gamma = \gamma_\theta \circ \gamma_\eta$, we have

$$G\left\{t|\, \boldsymbol{\theta}(\cdot), \boldsymbol{\eta}(\cdot)\right\} = \left\{\frac{\theta(\cdot)}{\theta(\cdot) + H(t)}\right\}^{\eta(\cdot)}. \tag{35}$$

Indeed,

$$\gamma_{\theta,\eta}(e^{-s}|\cdot) = \left[\frac{\theta(\cdot)}{\theta(\cdot) + s}\right]^{\eta(\cdot)}$$

is the Laplace transform of a Γ-distribution with scale parameter θ and shape parameter η, and we have the interpretation of the compound model (35) as a Γ–frailty model.

Note that since an exponentially distributed random variable corresponding to γ_θ is a continuous one, the above composition is not a particular case of (9).

The compound Θ is derived from the chain rule (20)

$$\Theta(x \mid \cdot, c) = \frac{\eta(\cdot) + c}{\theta(\cdot) - \log x}. \tag{36}$$

It is assumed that predictors depend on $\boldsymbol{\beta}, \boldsymbol{z}$ via the form $\beta_0 + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}$, where β_0 stands for the intercept term of the predictor. Also, different predictors have independent sets of regression coefficients $\theta = \theta(\beta_{\theta 0} + \boldsymbol{\beta}_\theta^{\mathrm{T}}\boldsymbol{z})$, $\eta = \eta(\beta_{\eta 0} + \boldsymbol{\beta}_\eta^{\mathrm{T}}\boldsymbol{z})$. To avoid overparameterization of the Γ-frailty model, the intercepts are fixed at zero. Based on the submodels, $\boldsymbol{\beta}_\theta$, $\boldsymbol{\beta}_\eta$ have the meaning of the log-odds- and log-hazards-ratio, respectively.

It should be noted that the direct formulation (35) of the Γ–frailty regression model by making shape and scale parameter into linear predictors offers certain advantages as compared to the traditional parameterization through the variance v of the frailty variable and restricted equal shape and scale parameters. With the traditional model $v = 0$ corresponding to the PH model is at the border of the parametric space $v \geq 0$. The traditional Γ frailty model is irregular for this reason. The model (35), on the contrary, is a regular one for any $\boldsymbol{\beta}$. We will confirm this observation in the simulation study by showing the validity of standard MLE theory for $\boldsymbol{\beta}$ with the model

(35) in Section 8.2. For the same reason we prefer (35) to the Dabrowska and Doksum model [Dabrowska and Doksum, 1988] that can be represented through a composition $\gamma = \gamma_{1/a} \circ \gamma_\theta \circ \gamma_a$, where γ_θ is an NTM–generating function for the PO model in the form (22), γ_a and $\gamma_{1/a}$ correspond to the PH model, and a is a scalar, independent of covariates

$$\gamma(x \,|\, \cdot) = \left\{ \frac{\theta(\cdot)x^a}{1 - \bar{\theta}(\cdot)x^a} \right\}^{\frac{1}{a}}, \quad a \geq 0. \tag{37}$$

The above model becomes the PO model in the form (22) if $a = 1$, and it becomes the PH model in the limit as $a \to 0$. With the above model, the PH assumption corresponds to the border of the parametric space ($a = 0$).

7 Identifiability

Given the combined model parameter $\omega = (\boldsymbol{\beta}, F)$, a naïve definition of identifiability would be

$$G(t \,|\, \omega_1, \boldsymbol{z}) \equiv_{t,\boldsymbol{z}} G(t \,|\, \omega_2, \boldsymbol{z}) \Rightarrow \omega_1 = \omega_2. \tag{38}$$

However, for semiparametric models, (38) does not hold. Even within the NTM class, presentation of a semiparametric model in terms of an NTM–generating function γ is not unique, as there is a number of ways to represent an arbitrary monotonic function. Indeed, expression (31) suggests that a transformation $p\{q(F)\}$, where p ia a tail function, q is an inverse tail function, and F is an arbitrary survival function, is again an arbitrary survival function. In other words, for any model γ, the family of NTM generating functions

$$\tilde{\gamma}(x \,|\, \cdot) = (\gamma \circ p \circ q)(x \,|\, \cdot) \tag{39}$$

represents the same semiparametric model for any p and q as defined above.

As an example, let us represent the two forms (22) and (23) as members of the family (39). Specifically, using (39), let p correspond to the logistic distribution, and q to the smallest extreme value distribution

$$p(x) = (1 + e^x)^{-1}, \qquad p^{-1}(x) = \log\{\text{Odds}(x)\},$$

$$q^{-1}(x) = \exp\{-\exp(x)\}, \quad q(x) = \log(-\log(x)).$$

Then, using the NTM–generating function (22) in conjunction with (39) and the functions p and q as defined above, we obtain

$$\tilde{\gamma}(x\,|\,\cdot) = \frac{\theta(\cdot)}{\theta(\cdot) - \log(x)}, \tag{40}$$

which is the NTM–generating function corresponding to the PO model in the form of exponential frailty model (23).

As a consequence of the above observations, frailty distributions are not identifiable unless the model is restricted. Such a restriction is provided, for example, by assuming G to be in the NTM class, and fixing the parametric form of the model–generating function γ. Then, if γ is an absolutely continuous distribution function on [0,1], then γ is a strictly increasing function, and F is identifiable, given $\boldsymbol{\beta}$, as $F = \gamma^{-1}(G\,|\cdot)$.

There is no universal chain rule as far as the "parametric" identifiability with respect to $\boldsymbol{\beta}$ is concerned. For example, while a composition of (Improper PH)$\circ$(Proper PH) models is identifiable as a PHPH model (Section 6.2.1), the reversed order (Proper PH)$\circ$(Improper PH) leads to an nonidentifiable Improper PH model of the form $\gamma(x\,|\cdot) = \exp\{-\theta\eta(1-F)\}$. Additional restrictions on the parametric part of the model may be necessary to ensure identifiability. For example, intercept term in the proper models is restricted to zero to eliminate its interaction with the unrestricted nonparametric baseline survival function F. For cure models built by composition of a non-cure and cure model-generating functions, F is restricted to be zero at the last failure [Taylor, 1995, Tsodikov, 2002], and the intercept term is removed from the non-cure submodel predictor. The intercept parameter in the predictor of the cure submodel codes for the baseline cure rate.

Identifiability of the PH frailty model received much attention in econometric literature [Ebbers and Ridder, 1982, Heckman and Singer, 1984, Heckman, 1991], primarily in the parametric setting. For semiparametric models, these results are valid under similar restrictions, and do not hold in the (38) form. Same observation applies to the perceived identifiability of shared frailty distribution from marginals in the bivariate case. Nonidentifiablity of frailties is yet another argument to consider modelling on the frailty-free NTM level.

The nonidentifiability aspect discussed in this section could be used to our advantage. The functions γ, p and q could be optimized within the family (39) to maximize the speed of convergence of QEM or to ensure its applicability. For example, QEM is not applicable to the PO model in the form (22) when $\theta > 1$ because the assumption of nondecreasing posterior risk function does not hold. At the same time, QEM is applicable to (23) for any θ.

8 Data analysis

8.1 Real data example

As an example, we use data from the National Cancer Institutes Surveillance Epidemiology and End Results (SEER) program. Using the publicly available SEER database, 39393 cases of primary prostate cancer diagnosed in Greater San Francisco between 1973 and 2000 were identified. Prostate cancer specific survival was analyzed by stage of the disease (localized/regional, 35230 patients, vs. distant, 4163 patients). For the definition of stages as well as for other details of the data we refer the reader to SEER documentation http://seer.cancer.gov/.

Two basic models PH (25) and PO (23), and two hierarchical compound models produced by compositions of PH and PO model generating functions, Γ–frailty model (35), and the PHPH cure model (32), were applied to fit the data. Stage of the disease was represented through two indicator dummy variables combined into a vector $\boldsymbol{z}$. Local/Regional stage was considered as a baseline group and the corresponding regression coefficient restricted to 0 for identifiability. Regression coefficient β for the distant stage codes for the difference in survival between the two stages expressed either as a log hazards or log odds ratio, dependent on the type of model generating function where it is used. The basic models have one predictor $\theta(\beta, z) = \exp(\beta z)$, where z=Indicator("Distant stage"). Compound models have two predictors, $\theta(\beta_\theta, z) = \exp(\beta_\theta z)$ and $\eta(\beta_\eta, z) = \exp(\beta_\eta z)$ coding two hazard ratios, long-term effect and short-term effect, respectively, in the PHPH cure model, and odds (θ) and hazard (η) ratios in the Γ-frailty model. In the latter model, odds and hazard ratio predictors have the interpretation of the scale and shape parameter of the frailty distribution, respectively. Regression coefficients in the PH model (β_θ) and the PH submodels of the PHPH (β_θ, β_η) and Γ-frailty models (β_η) measure the disadvantage of being in the distant stage relative to local/regional stage as a relative risk. Regression coefficient in the PO model (β_θ), and the one in the PO submodel of the Γ-frailty models measure the difference from an opposite point of relative odds of survival.

Since risk and odds of survival are opposites (high risk is bad, high survival is good), these coefficients are expected to be of opposite signs for in the PO and the PH model fitted to the same data.

Observed (Kaplan–Meier) and expected model–based estimates of the survival functions by group are shown in Figure 2.

Parameter estimates and confidence intervals are shown in Table 1.

Model	Parameter	Point–estimate	Confidence interval	p-Value
PH	β_θ	2.380	(2.328,2.432)	<0.001
PO	β_θ	-3.086	(-3.162,-3.011)	<0.001
PHPH	Improper PH: β_θ Proper PH: β_η	1.065 1.788	(0.923,1.207) (1.620,1.956)	<0.001 <0.001
Γ-frailty	PO: β_θ PH: β_η	-3.369 -0.179	(-3.580,-3.158) (-0.301,-0.057)	<0.001 <0.001

Table 1: Parameter estimation and hypothesis testing for prostate cancer data based on PH, PO, PHPH and Γ–frailty models. Negative β in the PO effect and positive β in the PH effect correspond to worse survival and vise versa.

Confidence intervals and hypotheses testing is based on the inverse of the observed profile information matrix

$$I_{pr} = -\frac{\partial^2 \ell_{pr}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}},$$

where the profile likelihood ℓ_{pr} is given by (5). Outlined in the Appendix A.4 are the main results that lead to exact computation of I_{pr}, [Tsodikov and Garibotti, 2005].

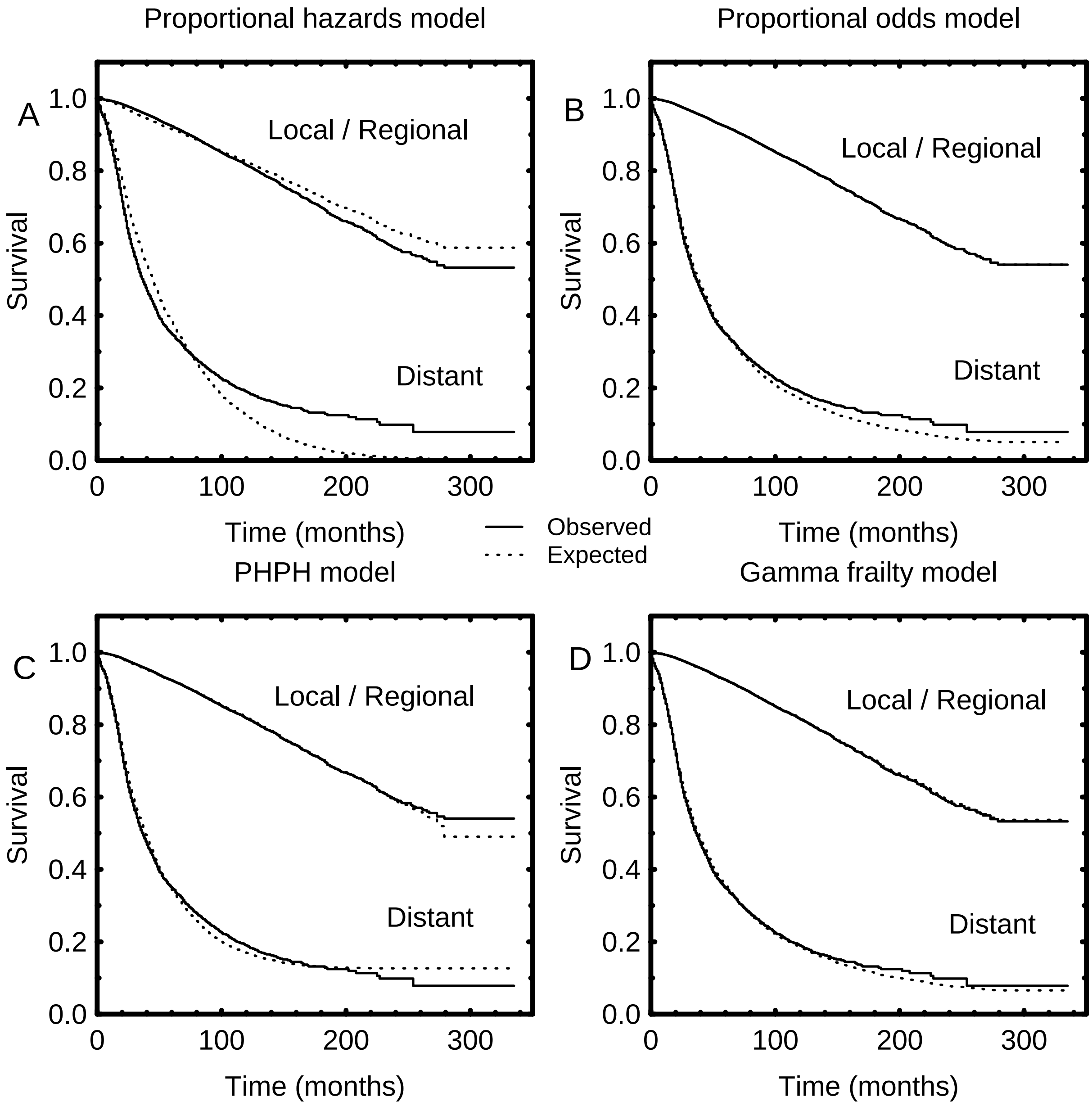


Figure 2: Prostate cancer cause-specific survival by stage. Observed (Kaplan-Meier) and expected survival curves for four models.

From Figure 2 it is evident that Γ-frailty model provides the best fit to the data. The PO model is second best. Given the hierarchical structure of Γ-frailty model, its goodness of fit can be tested vs. the PO model. This is a test for $\beta_\eta = 0$ in Γ-frailty model, and it results in a significant difference $\chi^2_1 = 7.50$, $p = 0.006$. The deviance with all other models exceeds 60, and we focus on the Γ-frailty model as the best choice at the level of model complexity considered so far. We could have tried to improve on the fit by using compositions of three or more submodels, but felt that the improvement over the Γ frailty model would be irrelevant for our data. All models indicate a highly significant effect of stage ($p < 0.0001$), which is a trivial conclusion in this case.

The validity of standard maximum likelihood theory as applied to the Γ-frailty model (35) will be studied by simulations in the next section. As the first observation, in Figure 3 we show that the form of the profile likelihood ℓ_{pr} in regression coefficients β_η (log hazards ratio) and β_θ (log odds ratio) is remarkably quadratic. In the next section we will verify by simulations that the curvature of the profile likelihood surface leads to consistent estimates of the standard errors of $\hat{\boldsymbol{\beta}}$.

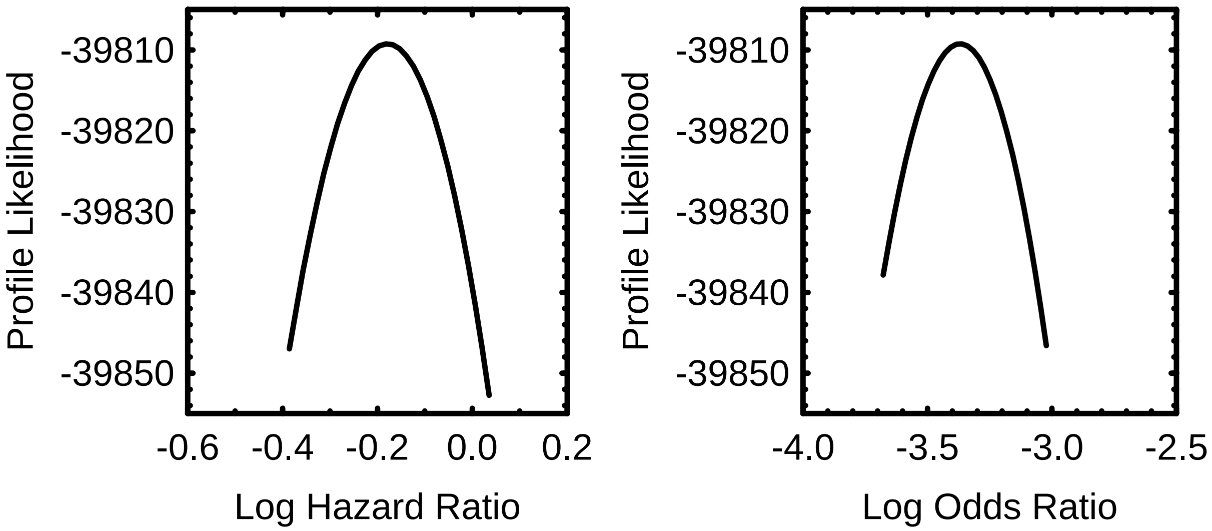


Figure 3: Profile likelihood as a function of regression coefficients sampled around the MLE point.

8.2 Simulations

We begin by fitting a parametric Γ-frailty model (35) to the prostate cancer data with the baseline survival function specified as Weibull distribution. The fit (not shown) is very similar to the semiparametric version of the model, and the parameter estimates are as follows, $\beta_\theta = -3.454$, $\beta_\eta = -0.215$, and [median of F]=265.571, [shape of F]=1.491.

Each simulation experiment was replicated 1000 times. Four sets of experiments were generated with samples sizes of 100 to 1000. Shown in Figure 4 are normal probability plots for the components of $\boldsymbol{\beta} = (\beta_\theta, \beta_\eta)^{\mathrm{T}}$. As evident from the figure, small sample size may be associated with some departure from normality of MLEs, however, with a sample size larger than 300 the estimates look perfectly normal. Shown in Table 2 are the results of simulations evaluating bias and variance of the estimates. Empirical means of $\hat{\boldsymbol{\beta}}$ show good correspondence to the true parameter values used to simulate the data and are within the margin of error expected from 1000 replicates. Empirical standard errors $\mathrm{S}_n\{\hat{\beta}\}$ estimated from replicated regression coefficients are in excellent correspondence with the $\mathrm{E}_n\{\hat{\sigma}_\beta\}$, the empirical mean of the replicated I_{pr}-based estimate of standard errors. The precision of variance estimation $\mathrm{S}_n\{\hat{\sigma}_\beta\}$ improves rapidly with the sample size.

9 Discussion

Recent years have seen an explosion of new survival models and model-specific estimation procedures. Mostly, new models are formulated on an ad-hoc basis and not much methodology is available to guide us on the model choice and appropriate estimation procedures. Although challenges of modern survival analysis will likely keep the business of model choice, estimation, identifiability and asymptotics largely on a case-by-case basis for some time, there is a continued effort to automate the process. This paper presents yet another step towards the goal by recognizing a mechanism of how models can be cloned while ensuring that the descendants are served computationally by a common Nonparametric Maximum Likelihood Estimation framework. We used frailty models and some associated compounding techniques as a motivation for the method of this paper and offered a way to build hierarchical families of semiparametric models that can be used to reproduce complex patterns of covariate effects using more than one predictor and to test model

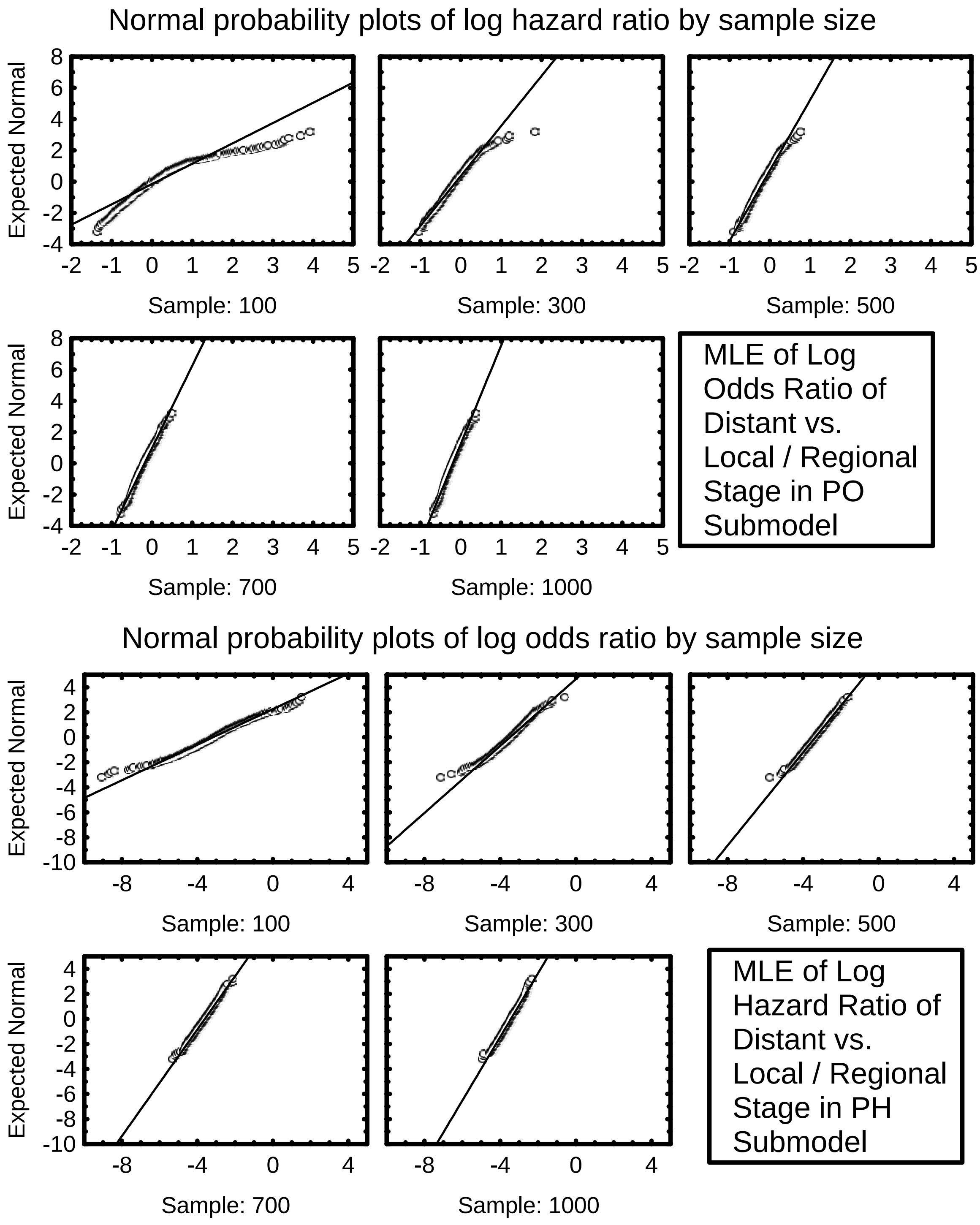


Figure 4: Normal probability plots

Parameter	$\mathrm{E}_n\{\hat{\beta}\}$	$\mathrm{S}_n\{\hat{\beta}\}$	$\mathrm{E}_n\{\hat{\sigma}_\beta\}$	$\mathrm{S}_n\{\hat{\sigma}_\beta\}$	Sample size
PH: β_θ	-3.168	1.394	1.451	0.415	100
PO: β_η	0.117	0.725	0.700	0.481	
PH: β_θ	-3.478	0.741	0.744	0.072	300
PO: β_η	-0.113	0.308	0.304	0.058	
PH: β_θ	-3.352	0.535	0.553	0.034	500
PO: β_η	-0.159	0.220	0.228	0.026	
PH: β_θ	-3.433	0.392	0.391	0.018	1000
PO: β_η	-0.197	0.158	0.157	0.012	

Table 2: The results of computer simulation to verify asymptotic properties of profile likelihood based MLEs.

assumptions.

Oftentimes when we discover that one of our standard models is wrong for the data, heterogeneity in the form of a random effect is introduced to model the departure from the basic model. This strategy is associated with building models on the level of missing data and it requires a large amount of analytical work to specify the algorithms. If our goal is to come up with a suitable model for the data rather than to build the model on mechanistic premises, the heterogeneity instrument is neither convenient nor necessary. Non-frailty framework such as NTM-QEM discussed in this paper offers streamlined model building and specification of inference algorithms with minimal analytic effort.

While rigorous evidence of identifiability, consistency and efficiency is still spotty, we will have to resort to simulations in studying the asymptotic properties of estimation procedures. Construction of a general computational and model building framework could stimulate targeted efforts to develop rigorous asymptotic theory for certain classes of models.

The non-identifiability aspect discussed in Section 7 is quite intriguing. A transformation (39) as well as other alternative parameterizations of the model generating function do not amount to a re-parameterization of the model and invariant MLEs. While regularity, asymptotics, convergence and other properties of inference procedures change under such transformations, the semiparametric model stays the same. This poses an interesting question of optimizing inference procedures over the functional class of model-invariant transformations.

Interesting issues for future research include extensions of the principles presented in this paper to multivariate survival models and time-dependent covariates.

Acknowledgement

This research is supported by National Cancer Institute grant U01 CA97414, and Department of Defence grant DAMD17-03-1-0034.

References

O.O. Aalen. Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, 2:951–972, 1992.

S. Bennett. Analysis of survival data by the proportional odds model. *Statistics in Medicine*, 2:273–277, 1983.

S.R. Cheng, L.J. Wei, and Z Ying. Analysis of transformation models with censored data. *Biometrika*, 82:835–845, 1995.

S.R. Cheng, L.J. Wei, and Z Ying. Predicting survival probabilities with semiparametric transformation models. *Journal of the American Statistical Association*, 92(437):227–235, 1997.

D. Clayton and J. Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society, Series A*, 148: 82–117, 1985a.

D. Clayton and J. Cuzick. The semiparametric pareto model for regression analysis of survival times. *Bull. Internat. Statist. Inst.*, 51:1–18, 1985b.

D.R. Cox. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, B*, 34:187–220, 1972.

D.M. Dabrowska and K.A. Doksum. Estimation and testing in a two-sample generalized odds-rate model. *Journal of the Americal Statistical Association*, 83:744–749, 1988.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

J. R. Dixon, M. R. Kosorok, and B. L. Lee. Functional inference in semiparametric models using the piggyback bootstrap. *Annals of the Institute of Statistical Mathematics*, 2005. In press, http://www.stat.fsu.edu/ dixon/boottr4.pdf.

K.H. Doksum and M. Gasko. On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review*, 58:243–252, 1990.

C. Ebbers and G. Ridder. True and spurious durational dependence: The identifiability of the proportional hazards model. *Review of Economic Studies*, 43:403–409, 1982.

W. Feller. *An Introduction to Probability Theory and its Applications.* John Wiley and Sons, New York, 1971.

J. Heckman and B. Singer. The identifiability of the proportional hazards model. *Review of Economic Studies*, 51:231–241, 1984.

J. Heckman. Identifying the hand of past: Distinguishing state dependence from heterogeneity. *The American Economic Review*, 81:75–79, 1991.

P. Hougaard. Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71(1):75–83, 1984.

J.P. Klein. Semiparametric estimation of random effects using the Cox model based on the EM algorithm. *Biometrics*, 48:795–806, 1992.

A.Y.C. Kuk and C.-H. Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, pages 531–541, 1992.

K. Lange, D.R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, pages 1–59, 2000.

T.A. Moger, O.O. Aalen, K. Heimdal, and H.K. Gjessing. Analysis of testicular cancer data using a frailty model with familial dependence. *Statistics in Medicine*, 23:617–632, 2004.

S.A. Murphy and A.W. Van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95:449–485, 2000.

G.G. Nielsen, R.D. Gill, P.K. Andersen, and T.I. Sorensen. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19:25–43, 1992.

J.M.G. Taylor. Semi-parametric estimation in failure time mixture models. *Biometrics*, 51:899–907, 1995.

A. Tsodikov and G. Garibotti. Profile information matrix for nonlinear transformation models. *Lifetime Data Analysis*, 2005. Submitted.

A. Tsodikov, J.G. Ibrahim, and A.Y. Yakovlev. Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association*, 98:1063–1078, 2003.

A. Tsodikov. A proportional hazards model taking account of long-term survivors. *Biometrics*, 54:1508–1516, 1998.

A. Tsodikov. Semiparametric models of long- and short-term survival: An application to the analysis of breast cancer survival in Utah by age and stage. *Statistics in Medicine*, 21:895–920, 2002.

A. Tsodikov. Semiparametric models: a generalized self-consistency approach. *Journal of the Royal Statistical Society, Series B*, 65:759–774, 2003.

J.W. Vaupel, K.G. Manton, and E. Stallard. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16:439–454, 1979.

J.T. Wassel and M.L. Moeschberger. A bivariate survival model with modified gamma frailty for assessing the impact of interventions. *Statistics in Medicine*, 12:241–248, 1993.

C.F.J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11:95–103, 1983.

A Appendix

A.1 Proof of Proposition 4.1. Properties of the surrogate of posterior risk.

Observe that

$$\frac{\partial}{\partial x}\Theta(x\,|\cdot,0) = \frac{\Theta(x\,|\cdot,0)}{x}\left[\Theta(x\,|\cdot,1) - \Theta(x\,|\cdot,0)\right] \geq 0,$$

as Θ is a non-decreasing function of x. Also, since γ is strictly increasing,

$$\Theta(x\,|\cdot,0) = \frac{\partial \log \gamma(x\,|\cdot)}{\partial x} > 0.$$

The above two expressions imply

$$\Theta(x\,|\cdot,1) \geq \Theta(x\,|\cdot,0).$$

This proves (15). Now, observe that for PH mixture models, $\Theta(x\,|\cdot,c)$ is the conditional expectation of frailty random variable U, given observed event. This interpretation of (15) gives (16), and

$$\mathrm{E}\left\{U\,|\,0,\cdot,0)\right\} = \gamma'(1\,|\,\cdot).$$

Since for PH mixture models γ is a p.g.f.,

$$\mathrm{E}\{U\,|\,\cdot\} = \gamma'(1\,|\,\cdot).$$

Combining the above two expressions gives (17).

A.2 Proof of Proposition 5.1 The class of mixture models is closed with respect to composition

By the Bernstein theorem (Feller [1971]), we need to prove that $\gamma(e^{-s}|\cdot) = (\gamma_\theta \circ \gamma_\eta)(e^{-s}|\cdot)$ is a completely monotonic function. Let $\psi_{\cdot}(s) = \gamma_{\cdot}(e^{-s}|\cdot)$. We have $\psi(s) = \psi_\theta\left\{-\log\psi_\eta(s)\right\}$. For any functions ξ and ζ, the composition $\xi \circ \zeta$ is completely monotonic if ξ is completely monotonic, $\zeta > 0$, and ζ' is completely monotonic. Applied to the functions ψ, this means that we have to prove that for any completely monotonic function $\psi(s) > 0$, the function

$f(s) = \{-\log\psi(s)\}'$ is completely monotonic. It can be proved by induction that

$$(-1)^n f^{(n)}(s) = \sum_{k=1}^{n+1} a_{nk}(-1)^k \frac{\psi^{(k)}(s)}{\psi^{n-k+2}(s)},$$

where $a_{01} = 1$, $a_{n+1,1} = a_{n1}$, $a_{n+1,k} = a_{nk}(n-k+2) + a_{n,k-1}$, $k = 2, \ldots, n+1$, $a_{n+1,n+2} = a_{n,n+1}$, $n = 0, 1, \cdots$. From the above equations it follows that $a_{nk} > 0$ for any n, k. Also, $\psi(s) > 0$, $s > 0$, and since ψ is completely monotonic, $(-1)^k\psi^{(k)}(s) \geq 0$. Therefore, $(-1)^n f^{(n)}(s) \geq 0$, $s > 0$. End of proof.

A.3 Proof of Proposition 5.2. Composition chain rule.

Proof of first statement is a straightforward exercise in differentiation of compound functions entering (3). Validity of second statement follows from (20) upon observation that all components of (20) are nondecreasing functions in x. End of proof.

A.4 Profile information matrix

Let the vector h represent a set of jumps of the baseline cumulative hazard H. Implicit differentiation of the profile likelihood yields the following expression for the profile information matrix

$$I_{pr} = I_{\beta\beta} + \hat{h}_\beta^{\mathrm{T}} I_{hh} \hat{h}_\beta + \hat{h}_\beta^{\mathrm{T}} I_{h\beta} + I_{h\beta}^{\mathrm{T}} \hat{h}_\beta, \tag{41}$$

where

$$\hat{h}_\beta = \left.\frac{\partial \hat{h}}{\partial \beta}\right|_{\beta=\hat{\beta}} \qquad \text{and} \qquad I_{ab} = -\left.\frac{\partial^2 \ell(\beta, h)}{\partial a \partial b^{\mathrm{T}}}\right|_{(\hat{\beta},\hat{h})}$$

with a and b equal to β or h.

Notice that I_{pr} has dimension $d \times d$, $d = \dim(\beta)$. Therefore only a small matrix needs to be inverted in order to get an estimator of the covariance matrix of regression coefficients.

The difficulty in (41) is that since $\hat{h}(\beta)$ is defined implicitly, so is the potentially large Jacobian matrix $\partial\hat{h}/\partial\beta$. Therefore, the Jacobian is generally unavailable in a closed form and its computation is the crux of the matter. It can be shown [Tsodikov and Garibotti, 2005] that $\partial\hat{h}/\partial\beta$ satisfies a system of linear equations with a special structure.

$$(R+D)\frac{\partial\hat{h}}{\partial\beta_k}=b^{(k)},$$

where D be the diagonal matrix with elements

$$d_m=\frac{D_m}{(\hat{h}_m)^2},\quad m=1,\ldots,d,$$

$R=(R_{ml})$ with $R_{ml}=\sum_{i=\max\{m,l\}}^{n}a_i$, where

$$a_i=\sum_{j\in\mathcal{C}_i\cup\mathcal{D}_i}Q(F_i\,|\,\beta,z_{ij},c_{ij}),\quad i=1,\ldots,n,$$

$\mathcal{C}_i$ and $\mathcal{D}_i$ is a set of censored observations and failures at the ith time point, respectively,

$$Q(x\,|\,\cdot,c)=-x\frac{\partial\Theta(x\,|\,\cdot,c)}{\partial x}=-(\Theta(x\,|\,\cdot,c)-c)(\Theta(x\,|\,\cdot,c+1)-\Theta(x\,|\,\cdot,c)),$$

and for $k=1,\ldots,d$,

$$b^{(k)}=\left(-\sum_{(i,j)\in\mathcal{R}_1}\frac{\partial\Theta(F_i\,|\,\beta,z_{ij},c_{ij})}{\partial\beta_k},\ldots,-\sum_{(i,j)\in\mathcal{R}_n}\frac{\partial\Theta(F_i\,|\,\beta,z_{ij},c_{ij})}{\partial\beta_k}\right)^{\mathrm{T}}.$$

For each $k=1,\ldots,n$ the vector $\partial\hat{h}/\beta_k$ can be obtained from the following Proposition [Tsodikov and Garibotti, 2005].

Proposition A.1 *Let D be an $n\times n$ diagonal matrix with diagonal elements $d_i\neq 0$, $i=1,\ldots,n$. Let $R=(R_{kl})$ be an $n\times n$ matrix, with $R_{kl}=\sum_{i=\max\{k,l\}}^{n}a_i$, where a_i, $i=1,\ldots n$ are real numbers. Let b be an n-dimensional vector.*

Define the functions $\varphi_k:\mathbb{R}\to\mathbb{R}$, $k=1,\ldots,n$ recursively as

$$\begin{aligned}\varphi_n(y)&=\frac{b_n}{d_n}-\frac{a_n}{d_n}y,\\ \varphi_k(y)&=\frac{1}{d_k}\left(b_k-\sum_{i=k}^{n}a_iy+\sum_{l=k+1}^{n}\sum_{i=k}^{l-1}a_i\varphi_l(y)\right),\quad k=n-1,\ldots,1,\end{aligned}$$

for y in $\mathbb{R}$. Let $\tilde{\varphi} : \mathbb{R} \to \mathbb{R}$ be the function given by $\tilde{\varphi}(y) = \sum_{k=1}^{n} \varphi_k(y)$ and let

$$\tilde{y} = \frac{\tilde{\varphi}(0)}{1 + \tilde{\varphi}(0) - \tilde{\varphi}(1)}.$$

Then the solution to the system of equations $(D + R)x = b$ is the n-dimensional vector $x = (\varphi_1(\tilde{y}), \ldots, \varphi_n(\tilde{y}))^T$.

Profile Information Matrix for Nonlinear Transformation Models

A. Tsodikov[†] and G. Garibotti[‡]

[†]University of California, Department of Public Health Sciences, Division of Biostatistics, One Shields Avenue, Davis, CA 95616, U.S.A., atsodikov@ucdavis.edu

[‡]Huntsman Cancer Institute, University of Utah, Division of Biostatistics, 2000 Circle of Hope, Salt Lake City, UT 84112, U.S.A, gilda.garibotti@hci.utah.edu.

Corresponding Author

ALEXANDER TSODIKOV, Ph.D.
DIVISION OF BIOSTATISTICS
DEPARTMENT OF PUBLIC HEALTH SCIENCES
UNIVERSITY OF CALIFORNIA
ONE SHIELDS AVENUE
DAVIS, CALIFORNIA 95616-8638
PHONE: (530) 752-0196
FAX: (530) 752-3239
Email: atsodikov@ucdavis.edu
http://phs.ucdavis.edu/

Abstract

For semiparametric models, interval estimation and hypotheses testing based on the information matrix for the full model is a challenge because of potentially unlimited dimension. Use of the profile information matrix for a small set of parameters of interest is an appealing alternative. Existing approaches for the estimation of the profile information matrix are either subject to the curse of dimensionality, or are ad-hoc and approximate and can be unstable and numerically inefficient. We propose a numerically stable and efficient algorithm that delivers exact observed profile information matrix for regression coefficients for the class of Nonlinear Transformation Models [Tsodikov, 2003]. The algorithm deals with the curse of dimensionality and requires neither large matrix inverses nor explicit expressions for the profile surface.

Keywords: profile likelihood, semiparametric models, information matrix

1 Introduction

In semiparametric models the parameter is partitioned as (β, H) with β a low-dimensional parameter of interest and H a high-dimensional nuisance parameter. For example, in semiparametric regression survival models, β is the vector of regression coefficients and H is the baseline cumulative hazard function estimated as a step-function by the Nonparametric Maximum Likelihood Estimator (NPMLE). The dimension of H is given by the number of distinct failure times and increases with the sample size.

Within the NPMLE framework the following tools are available for interval estimation and hypotheses testing for β.

1. *Likelihood Ratio.* The likelihood ratio statistic for testing $H_0 : \beta = \beta_0$ is defined as,

   $$\mathrm{LR}(\beta_0) = 2\left(\ell(\hat{\beta}, \hat{H}) - \ell(\beta_0, \hat{H}(\beta_0))\right),$$

   where ℓ is the log-likelihood function, $(\hat{\beta}, \hat{H})$ is the NPMLE of (β, H), and $\hat{H}(\beta)$ is the MLE of H given β. Although classical ML theory does not directly apply to unlimited dimension, for many semiparametric models LR has an asymptotic chi-square distribution with d degrees of freedom, where d is the dimension of β. A $(1-\alpha)\%$ confidence set for β is given by

   $$\left\{\beta :\ \mathrm{LR}(\beta) \leq C_{d,\alpha}\right\},$$

   where $C_{d,\alpha}$ is the α percentile of the chi-square distribution with d degrees of freedom. When the asymptotic distribution of LR is unknown, bootstrap can be used to approximate $C_{d,\alpha}$.

   The likelihood ratio approach for building confidence regions for β involves inverting the LR surface, which is quite computer intensive as repeated maximizations of the likelihood with respect to H are required.

2. *Wald Statistic.* An alternative method of inference for β is based on the Wald statistic defined as

   $$W(\beta) = (\hat{\beta} - \beta)^{\mathrm{T}} \Sigma_{\beta\beta}^{-1} (\hat{\beta} - \beta),$$

   where $\Sigma_{\beta\beta}$ is the β–submatrix of the inverse of the observed information matrix

   $$I = \left. \begin{pmatrix} -\frac{\partial^2 \ell(\beta,H)}{\partial\beta\partial\beta^{\mathrm{T}}} & -\frac{\partial^2 \ell(\beta,H)}{\partial\beta\partial H^{\mathrm{T}}} \\ -\frac{\partial^2 \ell(\beta,H)}{\partial H\partial\beta^{\mathrm{T}}} & -\frac{\partial^2 \ell(\beta,H)}{\partial H\partial H^{\mathrm{T}}} \end{pmatrix} \right|_{\beta=\hat{\beta}, H=\hat{H}} .$$

   Note that in the presence of nuisance parameters the information matrix needs to be inverted twice [Severini, 2000], p. 121, the first time in its high-dimensional full model form I, and the second time as a $\dim(\beta)$–submatrix of $\Sigma = I^{-1}$.

Under certain conditions, W is asymptotically equivalent to the likelihood ratio and has asymptotically a chi-square distribution with d degrees of freedom. In this case,

$$\{\beta:\ W(\beta) \leq C_{d,\alpha}\},$$

is a confidence set of approximate coverage probability $1-\alpha$.

The bottleneck of this procedure is the invertion of a potentially infinitely large matrix I.

The two methods of inference on β described above are based on the full model. An appealing alternative is to consider the so–called profile likelihood [Murphy and van der Vaart, 2000]

$$\ell_{pr}(\beta) = \max_{H} \ell(\beta, H).$$

The profile likelihood may be used as a likelihood for β. The MLE for β, the first component of the pair $(\hat{\beta}, \hat{H})$ that maximizes $\ell(\beta, H)$, is the maximizer of the profile likelihood function $\ell_{pr}(\beta)$.

Theoretical justification for the use of the profile likelihood for semiparametric models was given in [Murphy and van der Vaart, 2000, van der Vaart, 1998, Murphy and van der Vaart, 1997]. It was shown that profile likelihoods with nuisance parameter estimated out behave like ordinary likelihoods under regularity conditions. These conditions need to be verified on a case by case basis as the general theory remains a challenge. Theoretical justification has been obtained for the proportional odds (PO) model [Murphy and van der Vaart, 2000, Murphy et al., 1997] and the PH frailty models [Murphy, 1994, 1995, Parner, 1998, Kosorok et al., 2004].

The observed profile information matrix will be denoted I_{pr},

$$I_{pr} = -\left.\frac{\partial^2 \ell_{pr}(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}\right|_{\beta=\hat{\beta}}.$$

This matrix is asymptotically same as $\Sigma^{-1}_{\beta\beta}$, and summarizes partial information on β.

The Likelihood Ratio and Wald statistics based on ℓ_{pr} are easier to obtain than the ones based on the full model provided

- a numerically efficient method is available to profile out the nuisance parameter H, and

- it is possible to derive the exact observed profile information matrix or estimate it in a computationally efficient way.

Fulfilling both conditions is a challenge. First, maximization over H is a problem of potentially very large dimension. Second, in most cases ℓ_{pr} cannot be differentiated analytically. Several alternatives for estimating the profile information matrix have been proposed in the literature. However, they are all approximations, often difficult to calibrate in practice, and algorithms to obtain them are computationally costly.

In this paper we propose a computationally efficient exact solution for the class of semi-parametric Nonlinear Transformation Models (NTM) [Tsodikov, 2003]. The basic assumption that defines this model family is that the survival function at each timepoint t is a *function* of $H(t)$ mapping real numbers $[0, \infty] \to [0, 1]$ rather than a *functional* mapping a functional space to [0,1]. In other words, model-based survival function is obtained by plugging a cumulative hazard H or a baseline survival function $F = \exp(-H)$ into a suitably defined parametric function (so-called model-generating function, see Section 2). Note that a similar assumption underlies the von-Mises Calculus [van der Vaart, 1998], p.291. The NTM class includes the proportional hazards (PH) model, univariate PH frailty models], the proportional odds model, cure models such as the PHPH model [Tsodikov, 2002, Tsodikov et al., 2003]. A numerically efficient Quasi-EM algorithm, a subset of the MM family Lange et al. [2000] was developed to obtain the maximum profile likelihood for NTM models [Tsodikov, 2003]. The algorithm has since been used in computer intensive settings such as the bootstrap [Dixon et al., 2005].

The algorithm for the exact I_{pr} proposed in this paper works under the following two basic assumptions.

Independence of the future. Independence of the future means that the contribution to the likelihood of an observed event at time t depends on the past $H[0, t]$ of the function H, but not on the future.

Nonlinear Transformation Model Assumption. The survival function given covariates is specified as a parametric transformation of H. A detailed definition is given below.

We compare our method to the following three existing techniques used to estimate the profile information matrix that amount to particular forms of numerical differentiation of the second order.

1. *Discretized second derivative.* Corollary 3 of [Murphy and van der Vaart, 2000] shows that under certain conditions

$$-2\frac{\log \ell_{pr}(\hat{\beta} + h_n v_n) - \log \ell_{pr}(\hat{\beta})}{nh_n^2} \xrightarrow{P} v^{\mathrm{T}} I_{pr} v, \tag{1}$$

for all sequences $v_n \xrightarrow{P} v \in \mathbb{R}^d$ and $h_n \xrightarrow{P} 0$ such that $(\sqrt{n}h_n)^{-1} = O_P(1)$.

This result can be used to derive an estimate of I_{pr}. Note that this method requires careful maintenance of the speed of convergence of the sequence $\{h_n\}$ as the condition $(\sqrt{n}h_n)^{-1} = O_P(1)$ implies that the convergence should be neither too slow nor too fast. The reason is that the precision of discrete differential operator as $n \to \infty$ in the left side of (1) needs to be measured against the convergence of MLEs to the true value. Indeed, under regularity conditions, the asymptotic expansion of the likelihood ratio statistic about the MLE $\hat{\beta}$ has the form $n(\beta-\hat{\beta})^{\mathrm{T}} I_{pr}(\beta-\hat{\beta}) + o_P(\sqrt{n}\,\|\beta - \beta^*\| + 1)$, where β^* is the true value. The procedure (1) is designed to extract the quadratic term by setting $\beta = \hat{\beta} + h_n v_n$, and by ensuring the $1/\sqrt{n}$ rate of convergence of β to simultaneously $\hat{\beta}$ and β^* so that the quadratic term is indeed the dominant one. Otherwise, the expansion would be dominated by its $o_P(1)$ part if h_n is too fast or by $o_P(\sqrt{n}\,\|\beta - \beta^*\|)$ if h_n is too slow.

See Section 4 for further details and implementation of this method.

2. *Fitting a Quadratic Form.* Asymptotically, under regularity conditions, the profile likelihood surface around the true β is quadratic. Nielsen et al. [1992] proposed fitting a quadratic form to $\ell_{pr}(\beta)$ in some domain around the maximum likelihood estimator, $\hat{\beta}$, and to derive an approximate profile information matrix using the estimated coefficients of the form. Note that globally the likelihood surface is not quadratic. The quadratic approach is difficult to implement as a sufficiently small domain around $\hat{\beta}$ where the likelihood surface can be well approximated by a quadratic form is not well defined. Misspecification of this domain with the quadratic method often leads to estimates of the profile information matrix that are not positive definite, particularly if the number of covariates is large. Yet the domain needs to be large enough to ensure adequate precision and sufficient sample size representing the number of likelihood evaluations within the domain. This balancing act is notoriously difficult as the true variance is unknown and the likelihood surface is specific to the data set being analyzed.

3. *Numerical Differentiation of the Profile Likelihood.* Standard numerical algorithms can be used to numerically differentiate the profile likelihood function. We use Ridder's method [Press et al., 1994] in the examples presented in Section 4. The difficulties in the implementation of this idea are similar to the ones with the Quadratic Form approach. Numerical differentiation requires choosing a tolerance for the estimation of the derivatives, and typically involves interpolation of the function. The precision and speed of these methods are in inverse relashionship and they vary widely dependent on the tolerance. Since likelihood surface is dataset–specific, this method may require calibration and tuning for a particular dataset.

Approximating nature of the standard approaches outlined above, the need to balance various tradeoffs in their implementation, and a likely need to tweak implementation based on the dataset at hand, makes it difficult to develop these approaches to the point of automation sufficient for use in standard statistical software.

The algorithm proposed in this paper is exact, automatic and requires no tuning or calibration. This makes it an attractive alternative, particularly with statistical software applications in mind.

The PO model is used in this paper to compare via simulations the performance of the three estimation methods for I_{pr} and the proposed exact algorithm. For different sample sizes, the approaches were compared in terms of the number of operations required to achieve a reasonable standardized precision. Naturally, the exact method outperforms any approximating method if an ever better precision is demanded. In our numerical study we focus on practical precisions where approximating methods could nevertheless represent a viable competition to an exact procedure. Numerical efficiency and precision of the computation of I_{pr} is of great importance for variable selection procedures. In an example involving 7 variables, backward variable selection using the Wald statistic based on the exact profile information matrix took less than one third of the time of the quadratic approach. We also compared the estimation methods in terms of relative error. Of all approximating methods, the numerical approach has the smallest relative error.

As a result of these studies we believe that the exact method should be the primary choice for Nonlinear Transformation models.

2 Nonlinear Transformation Models

Nonlinear transformation models (NTM) are defined as follows [Tsodikov, 2002, 2003].

Definition 1 *Let $\gamma(x \mid \beta, z)$ be a parametrically specified distribution function with x–domain of $[0,1]$. Let $F(t)$ be a nonparametrically specified baseline survival function. A semiparametric regression survival model is called a Nonlinear Transformation Model if, conditional on the covariates z, its survival function G can be represented in the form*

$$G(t \mid \beta, z) = \gamma(F(t) \mid \beta, z). \tag{2}$$

The function γ is called the NTM-generating function.

Note that $F(t) = \exp(-H(t))$ where $H(t)$ is the baseline cumulative hazard function. With this in mind we can write the hazard function of the model as

$$\lambda(t \mid \beta, z) = \frac{\gamma'(F(t) \mid \beta, z)}{\gamma(F(t) \mid \beta, z)} F(t)h(t), \tag{3}$$

where $h(t) = H'(t)$ is the baseline hazard function.

In [Tsodikov, 2003] a Quasi-EM (QEM) point estimation algorithm for the NTM was developed and conditions that ensure its convergence were given.

The algorithm solves a functional self-consistency score equation of the form $H = \psi(\beta, H)$ for H, where ψ is a mapping that generalizes a Nelson-Aalen-Breslow estimator for the proportional hazards model so that its denominator depends on H as well as β. Functional iterations

$$H^{(k+1)} = \psi\left(\beta, H^{(k)}\right), \;\; k = 1, 2, \ldots \tag{4}$$

are exercised until $\hat{H}$, the fixed-point of ψ, has been approximated, $H^{(k)} \to \hat{H}$, as $k \to \infty$, see [Tsodikov, 2003] for details.

Although any parameterization of γ in terms of β and z is allowed, in the examples we assume that γ is parameterized through a set of parameters/predictors $\theta,\ \eta, \ldots$, where each predictor is further parameterized using generally different sets of regression coefficients $\beta_1, \beta_2, \ldots$, so that $\theta = \exp(\beta_1^{\mathrm{T}} z)$, $\eta = \exp(\beta_2^{\mathrm{T}} z)$, $\ldots$.

2.1 Profile Likelihood Approach

Let t_i, $i = 1, \ldots, n$ be a set of failure times, arranged in ascending order, $t_{n+1} := \infty$. Associated with each t_i is a set of subjects $\mathcal{D}_i$ with covariates $z_{ij},\ j \in \mathcal{D}_i$ who fail at t_i, and a set of subjects $\mathcal{C}_i$ with covariates $z_{ij},\ j \in \mathcal{C}_i$ who are censored at time $t \in [t_i, t_{i+1})$. The observed event for the subject ij is a triple (t_i, z_{ij}, c_{ij}), where c is a censoring indicator, $c = 1$ if failure, $c = 0$ if right censored. Let H be the baseline cumulative hazard, with $H(0) = 0$. We assume than $H(t)$ is a step function with jumps at the failure times t_i, $i = 1, \ldots, n$. As a step-function, H can be characterized by the vector $h = (h_1, ..., h_n)$, where $h_i = \Delta H_i$ is the jump of H at t_i. With this notation, under an NT model (2), (3) and non-informative censoring, the likelihood of survival data takes the form

$$\ell = \sum_{i=1}^{n} D_i \log(h_i) + \sum_{i=1}^{n} \sum_{j \in \mathcal{C}_i \cup \mathcal{D}_i} \log \vartheta(F_i \,|\, \beta, z_{ij}, c_{ij}),$$

where

$$\vartheta(x \,|\, \beta, z, c) = x^c \frac{\partial^c \gamma(x \,|\, \beta, z)}{\partial x^c},$$

$\partial^0 \gamma / \partial x^0 = \gamma$, D_i is the number of failures associated with t_i and

$$F_i = F(t_i) = \exp(-\sum_{l=1}^{i} h_l).$$

The profile likelihood is defined as a supremum of the full likelihood ℓ taken over the nonparametric part of the model

$$\ell_{pr}(\beta) = \max_h \ell(\beta, h).$$

The MLE of h for a given β will be denoted $\hat{h}(\beta) = (\hat{h}_1, \ldots, \hat{h}_n)$, with $\hat{h}_k = \hat{h}_k(\beta)$, then $\ell_{pr}(\beta) = \ell(\beta, \hat{h}(\beta))$.

Differentiating ℓ with respect to h and setting the score equal to 0 we obtain $\hat{h}(\beta)$ as the solution of the functional self-consistency equation

$$\hat{h}_m = \frac{D_m}{\sum_{(i,j)\in\mathcal{R}_m} \Theta(F_i \,|\, \beta, z_{ij}, c_{ij})}, \quad m = 1, \ldots, n, \tag{5}$$

where F_i is a function of $h_1, \ldots, h_i$,

$$\Theta(x \,|\, \beta, z, c) = -\frac{\partial \log \vartheta(x \,|\, \beta, z, c)}{\partial x} = c + x\frac{\gamma^{(c+1)}(x \,|\, \beta, z, c)}{\gamma^{(c)}(x \,|\, \beta, z, c)}, \tag{6}$$

and $\mathcal{R}_m$ is the set of subjects at risk just prior to t_m, $\mathcal{R}_m = \{(i,j) \,:\, i \geq m,\ j \in \mathcal{C}_i \cup \mathcal{D}_i\}$.

2.2 Point estimation

Point estimation proceeds along the lines of the following nested procedure,

- maximize $\ell_{pr}(\beta)$ by a conventional nonlinear programming method, for example, the Powell method [Press et al., 1994],
- for each β demanded in the above maximization procedure, find $\max_h \ell(\beta, h)$ as the fixed point of (5).

The Quasi-EM algorithm makes use of the straightforward recursion to obtain the profile likelihood,

$$h_m^{(k+1)} = \frac{D_m}{\sum_{(i,j)\in\mathcal{R}_m} \Theta(\exp(-\sum_{l=1}^{i} h_l^{(k)}) \,|\, \beta, z_{ij}, c_{ij})}, \ k = 1, 2, \ldots;\ m = 1, \ldots, n, \tag{7}$$

where k counts iterations. Note that an increment of k occurs only once all the parameters h_i, $i = 1, \ldots, n$ have been updated.

It can be shown that if Θ is nondecreasing in x, each update of H using (7) strictly improves the likelihood, given β. This guarantees convergence of the sequence of likelihood values $\ell(\beta, h^{(k)})$ to $\ell(\beta, \hat{h}(\beta))$ under fairly general conditions [Tsodikov, 2003]; that is, $\hat{h}(\beta)$ is the fixed point of the recursion given in (7).

It should be noted that the proposed information matrix algorithm is not contingent on using a specific method for point estimation. Yet it builds on the idea of the self-consistency through implicit differentiation of the self-consistency equation.

3 Profile Information Matrix

The profile information matrix is the observed information matrix derived from the profile likelihood,

$$I_{pr}(\beta) = -\frac{\partial^2 \ell_{pr}(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}.$$

Implicit differentiation of the profile likelihood yields the following expression for the profile information matrix

$$I_{pr}(\beta) = I_{\beta\beta} + h_\beta^{\mathrm{T}} I_{hh} h_\beta + h_\beta^{\mathrm{T}} I_{h\beta} + I_{h\beta}^{\mathrm{T}} h_\beta + \sum_{m=1}^{n} \ell_{h_m}\, h_{m,\beta\beta}, \tag{8}$$

where $h = h(\beta)$ is some function of β, and

$$h_\beta = \frac{\partial h(\beta)}{\partial\beta}, \quad I_{ab} = -\frac{\partial^2 \ell}{\partial a \partial b^{\mathrm{T}}}, \quad \ell_{h_m} = \frac{\partial \ell(\beta, h)}{\partial h_m}, \text{ and } h_{m,\beta\beta} = \frac{\partial^2 h_m(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}},$$

with a and b equal to β or h.

When evaluated at the MLE $\hat{h}(\beta)$, where $\hat{h}$ is a function defined implicitly as the solution of the score equation

$$\ell_h(\beta, h) = 0 \;\Rightarrow\; \hat{h} = \hat{h}(\beta), \tag{9}$$

the information matrix simplifies to

$$I_{pr} = I_{\beta\beta} + I_{\hat{h}\beta}^{\mathrm{T}} \hat{h}_\beta. \tag{10}$$

Indeed, by virtue of the score equation (9),

$$\ell_h(\beta, \hat{h}(\beta)) \equiv 0. \tag{11}$$

Differentiating (11) with respect to β, we also have

$$\frac{d\ell_h(\beta, \hat{h}(\beta))}{d\beta} = I_{\hat{h}\beta} + I_{\hat{h}\hat{h}} \hat{h}_\beta \equiv 0, \tag{12}$$

with (10) now following from (8) on substitution of (11) and (12).

It should be noted, however, that unless the score equation (9) is solved for h *exactly*, the short form of the observed profile information matrix (10) is generally not going to be symmetric. Except in the Cox model, there is no closed form solution to $\hat{h}$, and this function is an output of a numerical algorithm such as (7) converging to $\hat{h}$ with some tolerance. To preserve the symmetry of I_{pr}, we prefer to keep some of the theoretically redundant terms in (8) and use the form

$$I_{pr}(\beta) = I_{\beta\beta} + \hat{h}_\beta^{\mathrm{T}} I_{\hat{h}\hat{h}} \hat{h}_\beta + \hat{h}_\beta^{\mathrm{T}} I_{\hat{h}\beta} + I_{\hat{h}\beta}^{\mathrm{T}} \hat{h}_\beta. \tag{13}$$

Notice that I_{pr} has dimension $d \times d$, $d = \dim(\beta)$. Therefore only a small matrix needs to be inverted in order to get an estimator of the covariance matrix of regression coefficients.

The difficulty in (13) is that since $\hat{h}(\beta)$ is defined implicitly, so is the potentially large Jacobian matrix $\partial\hat{h}/\partial\beta$. Therefore, the Jacobian is generally unavailable in a closed form. The success in the calculation of the profile information matrix is determined by the existence of an efficient numerical method to compute $\partial\hat{h}/\partial\beta$. Generally, computation of $\partial\hat{h}/\partial\beta$ is as difficult as taking the inverse of the original full model information matrix ($O(n^3)$ operations required), and this derivation defeats the purpose. However, if the functional $\vartheta(H, t\,|\cdot)$ that defines model contributions to the likelihood depends on (H, t) only through $H(t)$, which is the case for the NT models (2), $\partial\hat{h}/\partial\beta$ can be obtained by solving a system of linear equations with a special structure. This specific structure of the linear system can be exploited to derive an efficient numerical solution given in Proposition 1.

We first show how to obtain $I_{\beta\beta}$, $I_{h\beta}$ and I_{hh}.

The H–score of an NT model is,

$$\frac{\partial\ell}{\partial h_k} = \frac{D_k}{h_k} - \sum_{(i,j)\in\mathcal{R}_m} \Theta(F_i\,|\,\beta, z_{ij}, c_{ij}).$$

Differentiating the H–score with respect to β we get,

$$-\frac{\partial\ell^2}{\partial h_k \partial\beta_m} = \sum_{(i,j)\in\mathcal{R}_k} \frac{\partial\Theta}{\partial\beta_m}(F_i\,|\,\beta, z_{ij}, c_{ij}).$$

Evaluation of derivatives of Θ or γ with respect to β depends on the parameterization of the model's predictor as a function of explanatory variables z, which is model–specific. Once a model is specified, the calculation of $I_{\beta\beta}$ and $I_{h\beta}$ is straightforward.

Since $F_i = \exp(-\sum_{l=1}^{i} h_l)$, we have

$$\frac{\partial\Theta(F_i\,|\,\cdot)}{\partial h_m} = \begin{cases} Q(F_i\,|\,\cdot), & m \le i, \\ 0, & m > i, \end{cases} \tag{14}$$

where

$$Q(x \mid \cdot, c) = -x \frac{\partial \Theta(x \mid \cdot, c)}{\partial x} = -(\Theta(x \mid \cdot, c) - c)(\Theta(x \mid \cdot, c+1) - \Theta(x \mid \cdot, c)), \tag{15}$$

and "'·"' stands for "'β, z"'. Note that $\partial \Theta(F_i \mid \cdot)/\partial h_m$ is a constant in m for $m \leq i$ or $m > i$.

From (14) it follows that,

$$-\frac{\partial^2 \ell}{\partial h_k \partial h_m} = \sum_{(i,j) \in \mathcal{R}_{\max\{k,m\}}} Q(F_i \mid \beta, z_{ij}, c_{ij}) + \frac{D_k}{h_k^2} 1_{\{k=m\}},$$

where

$$1_{\{k=m\}} = \begin{cases} 1, & k = m, \\ 0, & k \neq m. \end{cases}$$

From this we get I_{hh}.

Now we turn our attention to the Jacobian $\partial \hat{h}/\partial \beta$. Proposition 1 gives the main result used to efficiently calculate $\partial \hat{h}/\partial \beta$ in the case of NT models. Its proof is given in the Appendix.

Proposition 1 *Let D be an $n \times n$ diagonal matrix with diagonal elements $d_i \neq 0$, $i = 1, \ldots, n$. Let $R = (R_{kl})$ be an $n \times n$ matrix, with $R_{kl} = \sum_{i=\max\{k,l\}}^{n} a_i$, where a_i, $i = 1, \ldots n$ are real numbers. Let b be an n-dimensional vector.*

Define the functions $\varphi_k : \mathbb{R} \to \mathbb{R}$, $k = 1, \ldots, n$ recursively as

$$\begin{aligned} \varphi_n(y) &= \frac{b_n}{d_n} - \frac{a_n}{d_n} y, \\ \varphi_k(y) &= \frac{1}{d_k} \left(b_k - \sum_{i=k}^{n} a_i y + \sum_{l=k+1}^{n} \sum_{i=k}^{l-1} a_i \varphi_l(y) \right), \quad k = n-1, \ldots, 1, \end{aligned}$$

for y in $\mathbb{R}$. Let $\tilde{\varphi} : \mathbb{R} \to \mathbb{R}$ be the function given by $\tilde{\varphi}(y) = \sum_{k=1}^{n} \varphi_k(y)$ and let

$$\tilde{y} = \frac{\tilde{\varphi}(0)}{1 + \tilde{\varphi}(0) - \tilde{\varphi}(1)}.$$

Then the solution to the system of equations $(D + R)x = b$ is the n-dimensional vector $x = (\varphi_1(\tilde{y}), \ldots, \varphi_n(\tilde{y}))^T$.

We now show that the Jacobian $\partial\hat{h}/\partial\beta$ satisfies a relationship of the form as discussed in Proposition 1. Differentiating the self-consistency equation (5) implicitly, we get that $\hat{h}$ satisfies the relationship

$$\frac{\partial\hat{h}_m}{\partial\beta_k} = -\frac{\hat{h}_m^2}{D_m}\left(\sum_{l=1}^{n}\sum_{(i,j)\in\mathcal{R}_{\max\{m,l\}}} Q(F_i\,|\,\beta, z_{ij}, c_{ij})\frac{\partial\hat{h}_l}{\partial\beta_k} + \sum_{(i,j)\in\mathcal{R}_m}\frac{\partial\Theta}{\partial\beta_k}(F_i\,|\,\beta, z_{ij}, c_{ij})\right), \quad (16)$$

where Q is the function given in (15).

Let D be the diagonal matrix with elements

$$d_m = \frac{D_m}{(\hat{h}_m)^2}, \quad m = 1, \ldots, d.$$

Let $R = (R_{ml})$ with $R_{ml} = \sum_{i=\max\{m,l\}}^{n} a_i$, where

$$a_i = \sum_{j\in\mathcal{C}_i\cup\mathcal{D}_i} Q(F_i\,|\,\beta, z_{ij}, c_{ij}), \quad i = 1, \ldots, n$$

and for $k = 1, \ldots, d$ let

$$b^{(k)} = \left(-\sum_{(i,j)\in\mathcal{R}_1}\frac{\partial\Theta(F_i\,|\,\beta, z_{ij}, c_{ij})}{\partial\beta_k}, \ldots, -\sum_{(i,j)\in\mathcal{R}_n}\frac{\partial\Theta(F_i\,|\,\beta, z_{ij}, c_{ij})}{\partial\beta_k}\right)^{\mathrm{T}}.$$

It follows from (16) that

$$\frac{\partial\hat{h}}{\partial\beta_k} = -D^{-1}\left(R\frac{\partial\hat{h}}{\partial\beta_k} - b^{(k)}\right).$$

Hence,

$$(R + D)\frac{\partial\hat{h}}{\partial\beta_k} = b^{(k)}.$$

Therefore, for each $k = 1, \ldots, n$ the vector $\partial\hat{h}/\beta_k$ can be obtained from Proposition 1. We now have all the components of (13) defined. This completes the exposition of our method.

4 Examples

In the examples we compare the performance of four methods to compute the observed profile information matrix. A brief explanation of the methods and details on how they were implemented in our examples are given below.

1. *Discretized.* The estimation is based on the result of Corollary 3 in Murphy and van der Vaart [2000]. Under certain conditions

   $$-2\frac{\log \ell_{pr}(\hat{\beta}+h_n v_n)-\log \ell_{pr}(\hat{\beta})}{nh_n^2} \xrightarrow{P} v^{\mathrm{T}} I_{pr} v, \tag{17}$$

   for all sequences $v_n \xrightarrow{P} v \in \mathbb{R}^d$ and $h_n \xrightarrow{P} 0$ such that $(\sqrt{n}h_n)^{-1} = O_P(1)$.

   In order to estimate all the elements of I_{pr}, we chose $v = e_i$, $i = 1, \ldots, d$ and $v = e_i + e_j$, $1 \leq i < j \leq d$, where e_i are Euclidean basis vectors.

   We set $v_n \equiv v$, and $h_n = 10/(\sqrt{n}C^k)$ with $C = 1.4$ and k such that $|f(10/(\sqrt{n}C^k)) - f(10/(\sqrt{n}C^{k-1}))| < 0.001$, where $f(h)$ is the left hand side of equation (17) considered as a function of h. This procedure was motivated by Dixon et al. [2005] who considered a choice of h_n in the one dimensional ($d = 1$) situation.

2. *Quadratic.* This approach approximates the profile likelihood surface by a quadratic form and derives the estimate of the information matrix from the coefficients of the form fitted to the surface. Specifically, let $\Delta\beta$ be a vector of deviations of the β values sampled in the vicinity of $\hat{\beta}$, and let $\Delta\ell_{pr}$ be the induced vector of deviations of the profile likelihood from its maximum value, $\ell_{pr}(\hat{\beta})$. Then, if $\Delta\beta$ is sufficiently small

   $$\Delta\ell_{pr} \approx \frac{1}{2}\Delta\beta^{\mathrm{T}} I_{pr}\Delta\beta.$$

   Fitting the quadratic form $(1/2)\Delta\beta^{\mathrm{T}} A\Delta\beta$ to points $(\Delta\beta, \Delta\ell_{pr})$ by least squares produces an estimate, $\hat{A}$, of the profile information matrix I_{pr}.

   In our implementation of this method we limit the domain to points that are not rejected at 0.05 significance level by the LR test (applied informally and disregarding the multi-comparison issue). In other words, points β are included if $-2\left\{\ell_{pr}(\hat{\beta}) - \ell_{pr}(\beta)\right\} \leq C_{d,0.05}$, where $C_{d,0.05}$ is the 0.05th upper tail percentile of the χ^2 distribution with $d = \dim(\beta)$ degrees of freedom. Since the validity of the quadratic approximation is itself a prerequisite for the validity of the likelihood ratio statistic, this choice is far from perfect. Yet this procedure would ensure a desired property of the domain shrinking with sample size, and we know of no better alternative.

3. *Numerical.* The calculation of the observed profile information matrix is carried on using Ridder's numerical differentiation of the profile likelihood function, see Press et al. [1994].

Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function. By definition, the derivative of f is the limit as $h \to 0$ of the incremental quotient

$$q(h) = \frac{f(x+h) - f(x)}{h}.$$

The basic idea of Ridder's method is to calculate $q(h)$ for several values of h, and then extrapolate the result to the limit $h = 0$. In the case of a function with domain on $\mathbb{R}^d$, the vector of first derivatives is obtained applying the algorithm on each coordinate at a time and leaving the other coordinates fixed.

Numerical experimentation showed that this approach gives the second derivatives of $\ell_{pr}(\beta)$ with very high precision, albeit at a greater computational cost than other methods.

4. *Exact.* This is the method developed in Section 3 of this paper for computation of the exact observed profile information matrix for NTM.

PO model will be used as a basis for all our comparisons. The validity of NPMLE and the profile likelihood for this model has been demonstrated elsewhere.

4.1 The Proportional Odds Model

Given covariates z, the survival function $G(t \,|\, \beta, z)$ of a PO model can be written in the form,

$$G(t \,|\, \beta, z) = G(t \,|\, \theta(\beta, z)) = \frac{\theta(\beta, z)}{\theta(\beta, z) + H(t)}, \tag{18}$$

where H is some nonparametrically specified baseline cumulative hazard function, and θ is a predictor. Since $H = -\log F$, the NTM–generating function of the PO model is

$$\gamma(x \,|\, \cdot) = \frac{\theta(\cdot)}{\theta(\cdot) - \log x}.$$

A characteristic feature of the PO model is that for any two values, θ_1, θ_2, of the predictor, the odds ratio

$$\frac{\text{Odds}(G(t \,|\, \theta_1))}{\text{Odds}(G(t \,|\, \theta_2))} = \frac{\theta_1}{\theta_2}$$

is constant in t.

It follows that

$$\Theta(x \,|\, \cdot, c) = \frac{c+1}{\theta(\cdot) - \log x}.$$

We consider an exponential parameterization of the predictor $\theta(\beta, z) = \exp(\beta^{\mathrm{T}} z)$. With this parameterization,

$$\frac{\partial\theta}{\partial\beta} = \theta z, \qquad \frac{\partial^2\theta}{\partial\beta\partial\beta^{\mathrm{T}}} = \theta z z^{\mathrm{T}}.$$

The following derivatives of Θ are necessary to specify the algorithm of Section 3,

$$\frac{\partial\Theta}{\partial\beta} = \frac{\partial\Theta}{\partial\theta}\theta z, \qquad \frac{\partial^2\Theta}{\partial\beta\partial\beta^{\mathrm{T}}} = \left(\frac{\partial^2\Theta}{\partial\theta^2}\theta^2 + \frac{\partial\Theta}{\partial\theta}\theta\right) z z^{\mathrm{T}},$$

where

$$\frac{\partial\Theta(x\,|\,\cdot, c)}{\partial\theta} = -\frac{c+1}{(\theta(\cdot) - \log x)^2}, \quad \text{and} \quad \frac{\partial^2\Theta(x\,|\,\cdot, c)}{\partial\theta^2} = \frac{2(c+1)}{(\theta(\cdot) - \log x)^3}.$$

4.2 Real Data

As an example, we use data from the National Cancer Institutes Surveillance Epidemiology and End Results (SEER) program. Using the publicly available SEER database, 11621 cases of primary prostate cancer diagnosed in the state of Utah between 1988 and 1999 were identified. The following selection criteria were applied to a total of 19819 Utah–cases registered in the database: valid positive survival time, valid stage of the disease, age $\geq$ 18 years. Prostate cancer specific survival was analyzed by stage of the disease (localized/regional vs. distant). For the definition of stages as well as for other details of the data we refer the reader to SEER documentation http://seer.cancer.gov/.

The data analysis presented in this paper is a continuation of the one given in [Tsodikov, 2003]. Two groups of patients representing stage at diagnosis of the disease are considered, hence the predictor in the PO model has a single parameter β. The log odds ratio β measures the disadvantage of being in the distant stage relative to local/regional stage. The QEM algorithm was applied to fit the PO model to the data. The maximum likelihood estimate of β is $\hat{\beta} = -3.251$. Confidence intervals for β were obtained using the Wald statistic based on the profile information matrix. The confidence interval based on the quadratic approximation of the profile information matrix is $(-3.416, -3.086)$ and the one obtained through the exact profile information matrix is $(-3.415, -3.086)$. Excellent concordance of the two confidence intervals is due to the large sample size and the small dimension of the regression parameter, a situation when approximating methods tend to be accurate.

In the case of a single parameter, the observed profile information matrix is a scalar. The estimates of the observed profile information matrix are 142.1011, 141.2158 and 141.7424 for the Discretized, Quadratic and Numerical approaches respectively and the Exact value is 141.7423. Although the values are quite similar it is clear that the discretized and quadratic approaches depart from the true value.

4.3 Simulations

4.3.1 Simulations setup

We simulated age at diagnosis for an adult-onset disease using a proportional odds model. The example in its baseline survival is loosely based on indicence of prostate cancer. The baseline survival function was assumed to follow a Weibull distribution with the median of 38 years and shape of 1.8, the risk starting at the age of 18 (a fixed number of 18 years was added to survival time and censoring). With these parameters incidence before the age of 40 is negligible. Independent censoring mechanism was assumed. Censoring times were generated using Weibull distribution with a median of 46 years and shape parameter of 4. Observations in excess of 105 years were type-I censored at 105. Two covariates were introcuded, one categorical with 3 levels, and one continuous (a risk factor) with a range between -1 and 1. Values for both covariates were generated independently. The continuous covariate followed a uniform distribution. The discrete distribution for categorical covariate assumed the following probabilities, 0.7 (level 1), 0.5 (level 2), and 0.1 (level 3). The following covariate effects were assumed. An effect of the log odds ratio of 2 was assumed for a unit change in the continuous factor. Categorical covariate was assumed to have a progressing effect on the risk of the disease. The log odds ratios comparing level 2 and level 3 to the baseline level 1 were 1.5 and 2.5, respectively.

4.3.2 Speed

To assess the speed of performance of the four methods we calculated the number of operations required to compute the exact information matrix and its approximations. Evaluation of Θ, γ, their analytically specified derivatives or similar comparable procedures were counted as one operation. Figure 1 shows the number of operations by sample size and method. In order to make the performance results comparable, the precision of estimation algorithms was calibrated on an ad-hoc basis so that the relative error of the three methods (Discretized, Quadratic, Numerical) was approximately the same (0.02). For any method A, the relative error was defined as $\|I_{pr}(A) - I_{pr}(\text{Exact})\| \, / \, \|I_{pr}(\text{Exact})\|$, where $I_{pr}(A)$ is an estimate of the observed profile information matrix computed using method A, and the norm is defined as the sum of absolute values of all elements of the matrix. Regardless of the sample size, the exact calculation outperformed the approximate methods. Inference based on the discretized second derivative requires between 10 and 30 times as many operations as the exact calculation. The quadratic approach requires between 60 and 200 times as many operations as the calculation of the exact I_{pr} matrix. The numerical method is computationally very costly requiring between 600 and 7000 as many operations as the exact approach. However, the numerical approach behaves better than the other two methods in terms of relative error as

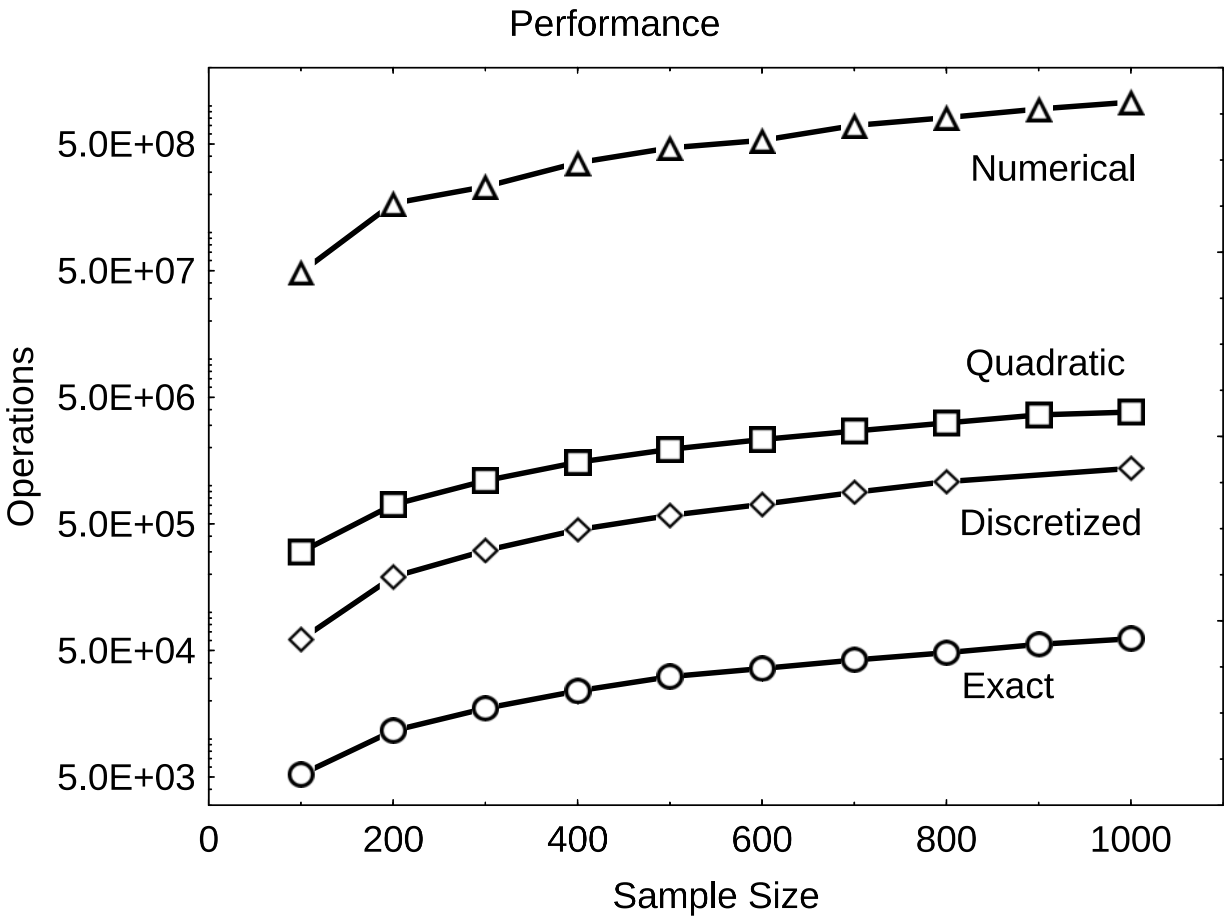


Figure 1: Operations by sample size characteristics of four methods of computation of the observed profile information matrix. The Exact method developed in this paper shows the highest numerical efficiency.

shown in Section 4.3.3.

4.3.3 Precision

A sample of size 500 was used to find the smallest possible relative error of the method when adjusting the different parameters involved on an ad-hoc basis. The best relative error achieved by the Discretized method was 0.01 and 8.13 10^5 operations were required. This number was 0.013 for the quadratic approach with 5.32 10^6 operations required, while the numerical approach achieved a relative error of 8 10^{-7} and required 3.87 10^8 operations. This example shows that the numerical approach is perhaps the only one of the approximating methods that can compete with the exact procedure in terms of precision required in real-life analysis. It's high computational cost though makes it a poor choice for variable selection and other procedures requiring repeated evaluations of I_{pr}.

4.3.4 Statistical properties

Three sets of experiments were performed with samples of size 100, 500, and 1000. For each sample size, 1000 simulated samples were generated. The covariance matrices based on I_{pr} were computed for each sample using the four approaches discussed. The mean and standard deviation of each of the entries of the estimators of covariance matrices under study were estimated from the 1000 replicates. In addition, point semiparametric MLE estimates of the three parameters entering profile likelihood (log odds ratios for the continuous factor and level 2 vs. 1 and 3 vs. 1 contrasts) were used to compute the empirical covariance matrix based on 1000 replicates. A comparison of the estimated means of the entries of the covariance matrices calculated using exact and numerical approaches with the empirical ones were used to evaluate how well these methods estimate the true finite sample variance-covariance. The results are shown in Table 1. Two factors are contributing to the distance between the exact and numerical approaches and the empirical one: the finite-sample bias of covariance estimates based on I_{pr}, and the bias in the estimate of I_{pr} by an approximating method (this latter bias does not pertain to the exact method). The following basic conclusions are evident from the Table 1.

1. All methods are much better at estimating variances (left half of the table) than covariances (right half of the table);

2. The precision of estimation of covariance improves rapidly with sample size;

3. Under all sample sizes the numerical approach showed excellent concordance with the exact method. This is in agreement with our earlier observation that of all approximat-

ing methods, the numerical approach is most precise albeit computationally costly. The quadratic method showed itself as least precise in this analysis;

4. For reasonably large sample sizes all methods are in good agreement with the empirical estimate.

5 Discussion

In this paper we have proposed a method to compute profile information matrix based on implicit differentiation of the self-consistency equation. Computationally the method outperformed all existing approaches to the best of our knowledge. An attractive property of the procedure is that it is exact contingent upon point estimates. Even though exact point estimates are hardly ever available, the precision of variance-covariance estimation is improved as the method does not add any error to the one associated with impresision of point estimates. Numerically efficient and stable procedures for point estimates have been developped earlier and provide a good complement to this methodology. We recommend the Exact method as a preferred choice with Nonlinear Transformation Models.

Since derivatives of the profile likelihood are defined implicitly, applying Newton-Raphson method to the profile likelihood for point estimation is a challenge. The Newton Raphson typically requires exact inverse Hessian matrix and is not guaranteed to converge if this matrix is approximated. The results of this paper can be used to provide an exact inverse Hessian matrix. at any point in the parameter space and thus enable the Newton Raphson method for use with the profile likelihood.

Acknowledgement

This research is supported by National Cancer Institute grant U01 CA97414, and Department of Defence grant DAMD17-03-1-0034.

	Mean (sd)					
method	σ_{11}	σ_{22}	σ_{33}	σ_{12}	σ_{13}	σ_{23}
			Sample size: 100			
empirical	0.405	0.597	0.207	0.102	0.044	0.045
exact	0.426 (0.170)	1.020 (0.347)	0.192 (0.041)	0.080 (0.022)	0.045 (0.035)	0.062 (0.046)
num	0.426 (0.170)	1.019 (0.346)	0.192 (0.041)	0.080 (0.022)	0.044 (0.035)	0.062 (0.046)
quad	1.009 (21.986)	0.103 (21.374)	1.139 (31.762)	-0.087 (9.674)	-0.188 (5.822)	-0.364 (7.674)
Mur	0.464 (0.203)	1.155 (0.411)	0.206 (0.046)	0.077 (0.025)	0.046 (0.042)	0.067 (0.054)
			Sample size: 500			
empirical	0.082	0.259	0.037	0.014	0.011	0.011
exact	0.074 (0.010)	0.245 (0.117)	0.036 (0.003)	0.016 (0.002)	0.009 (0.003)	0.013 (0.004)
num	0.074 (0.010)	0.245 (0.117)	0.036 (0.003)	0.016 (0.002)	0.009 (0.003)	0.013 (0.004)
quad	0.074 (0.034)	0.347 (0.490)	0.040 (0.039)	-0.003 (0.067)	0.002 (0.022)	0.032 (0.053)
Mur	0.076 (0.011)	0.260 (0.128)	0.036 (0.003)	0.014 (0.002)	0.008 (0.003)	0.013 (0.004)
			Sample size: 1000			
emp	0.037	0.116	0.018	0.006	0.005	0.006
exact	0.037 (0.003)	0.112 (0.028)	0.018 (0.001)	0.008 (0.001)	0.004 (0.001)	0.007 (0.001)
num	0.037 (0.003)	0.112 (0.028)	0.018 (0.001)	0.008 (0.001)	0.004 (0.001)	0.007 (0.001)
quad	0.038 (0.066)	0.114 (0.052)	0.018 (0.020)	0.002 (0.044)	0.003 (0.036)	0.008 (0.024)
Mur	0.037 (0.004)	0.116 (0.030)	0.018 (0.001)	0.007 (0.001)	0.004 (0.001)	0.006 (0.001)

Table 1: Empirical covariance matrix based on 1000 replicates and mean covariance matrix and sd based on 1000 replicates for the four methods studied

References

J.R. Dixon, M.R Kosorok, and B.L. Lee. Functional inference in semiparametric models using the piggyback bootstrap. *Annals of the Institute of Statistical Mathematics*, 57: 255–277, 2005.

M.R. Kosorok, B.L. Lee, and J.P. Fine. Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics*, pages 1448–1491, 2004.

K. Lange, D.R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, pages 1–59, 2000.

S.A. Murphy. Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, 22(2):712–731, 1994.

S.A. Murphy. Asymptotic theory for the frailty model. *The Annals of Statistics*, 23(1): 182–198, 1995.

S.A. Murphy and A.W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95:449–485, 2000.

S.A. Murphy and A.W. van der Vaart. Semiparametric likelihood ratio inference. *The Annals of Statistics*, 25:1471–1509, 1997.

S.A Murphy, A.J. Rossini, and A.W. van der Vaart. Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439): 968–976, 1997.

G.G. Nielsen, R.D. Gill, P.K. Andersen, and T.I. Sorensen. A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, 19: 25–43, 1992.

E. Parner. Asymptotoc theory for the correlated Gamma-frailty model. *The Annals of Statistics*, 26:183–214, 1998.

W.H. Press, B.P. Flannery, S.A. Teukolsky, and W.T. Vetterling. *Numerical Recipies in Pascal. The Art of Scientific Computing.* Cambridge University Press, New York, NY, 1994.

T.A. Severini. *Likelihood Methods in Statistics.* Oxford University Press, New York, NY, 2000.

A. Tsodikov. Semiparametric models: a generalized self-consistency approach. *Journal of the Royal Statistical Society, Series B*, 65:759–774, 2003.

A. Tsodikov. Semiparametric models of long- and short-term survival: An application to the analysis of breast cancer survival in Utah by age and stage. *Statistics in Medicine*, 21: 895–920, 2002.

A. Tsodikov, J.G. Ibrahim, and A.Y. Yakovlev. Estimating cure rates from survival data: An alternative to two-component mixture models. *Journal of the American Statistical Association*, 98:1063–1078, 2003.

A.W. van der Vaart. *Asymptotic statistics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, UK, 1998.

6 Appendix

Proof of Proposition 1

The equation $(D+R)x = b$ implies

$$d_k x_k + \sum_{l=1}^{n} R_{kl} x_l = b_k, \quad k = 1, \ldots, n. \tag{19}$$

Since $R_{kl} = \sum_{i=\max\{k,l\}}^{n} a_i$, it follows from (19) that for $k = 1, \ldots, n$,

$$\begin{aligned} b_k &= d_k x_k + \sum_{i=k}^{n} a_i \sum_{l=1}^{k} x_l + \sum_{l=k+1}^{n} \sum_{i=l}^{n} a_i x_l \\ &= d_k x_k + \sum_{i=k}^{n} a_i \sum_{l=1}^{k} x_l + \sum_{i=k+1}^{n} \sum_{l=k+1}^{i} a_i x_l \\ &= d_k x_k + \sum_{i=k}^{n} a_i \left(\sum_{l=1}^{n} x_l - \sum_{l=i+1}^{n} x_l \right). \end{aligned}$$

The second equality above is a consequence of a change of summation order.

Hence, solving the system of equations $(D+R)x = b$ is equivalent to solving the system

$$\begin{aligned} x_k &= \frac{1}{d_k}\left(b_k - \sum_{i=k}^{n} a_i y + \sum_{l=k+1}^{n} \sum_{i=k}^{l-1} a_i x_l \right), \quad k = 1, \ldots, n \\ y &= \sum_{l=1}^{n} x_l. \end{aligned}$$

Notice that $\{x_k\}$ are in fact functions of y obtained recursively, $x_k = \varphi_k(y)$, and $y = \sum_{k=1}^{n} \varphi_k(y) = \tilde{\varphi}(y)$. The solution of this system of equations is the vector $(\varphi_1(\tilde{y}), \ldots, \varphi_n(\tilde{y}))^{\mathrm{T}}$ where $\tilde{y}$ satisfies $\tilde{\varphi}(\tilde{y}) = \tilde{y}$. Since φ_k, $k = 1, \ldots, n$ are linear functions of y, so is $\tilde{\varphi}$. Hence, $\tilde{\varphi}(y) - y = ay + b$ with

$$a = \tilde{\varphi}(1) - 1 - \tilde{\varphi}(0) \quad \text{and} \quad b = \tilde{\varphi}(0).$$

Therefore,

$$\tilde{y} = \frac{\tilde{\varphi}(0)}{1 + \tilde{\varphi}(0) - \tilde{\varphi}(1)}.$$

End of proof.

STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2257

A population model of prostate cancer incidence

A. Tsodikov[1,*,†], A. Szabo[2,‡] and J. Wegelin[1]

[1] *Department of Public Health Sciences, Division of Biostatistics, University of California Davis, One Shields Avenue, Davis, CA 95616, U.S.A.*
[2] *Huntsman Cancer Institute at the University of Utah, 2000 Circle of Hope, Salt Lake City, UT 84112, U.S.A.*

SUMMARY

Introduction of screening for prostate cancer using the prostate-specific antigen (PSA) marker of the disease led to remarkable dynamics of the incidence of the disease observed in the last two decades. A statistical model is used to provide a link between dissemination of PSA and the observed transient population responses. The model is used to estimate lead time, overdiagnosis and other relevant characteristics of prostate cancer screening. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: screening; mixture models; prostate cancer

1. INTRODUCTION

Continuing controversy and discussions surround the issue of whether prostate-specific antigen (PSA) screening of asymptomatic men can be linked to recent decline in prostate cancer mortality [1–4]. Shown in Figure 1 is the age-adjusted incidence and cause-specific mortality curve as estimated from the Surveillance, Epidemiology and End Results (SEER) [5] database developed and maintained by the National Cancer Institute. While the incidence curve shows a sharp peak with the introduction of PSA testing in the late 1980s, mortality showed a much less dramatic behaviour. Advocates of screening argue that screening induces a favourable shift in the distribution of stage of the disease at diagnosis and that earlier detection and treatment should lead to better prognosis and reduce mortality from the disease. The difficulty in proving the point is rooted in the complexity of the changes that screening for a disease with high latent prevalence brings to the observed population statistics. Lead time and overdiagnosis are

[*] Correspondence to: Alex Tsodikov, Department of Public Health Sciences, Division of Biostatistics, University of California Davis, One Shields Avenue, Davis, CA 95616, U.S.A.
[†] E-mail: atsodikov@ucdavis.edu
[‡] E-mail: aszabo@hci.utah.edu

Contract/grant sponsor: National Cancer Institute Cancer Intervention and Surveillance Modelling Network; contract/grant number: U01 CA97414
Contract/grant sponsor: Department of Defense; contract/grant number: DAMD 7-03-1-0034

Received 20 October 2004
Accepted 1 April 2005

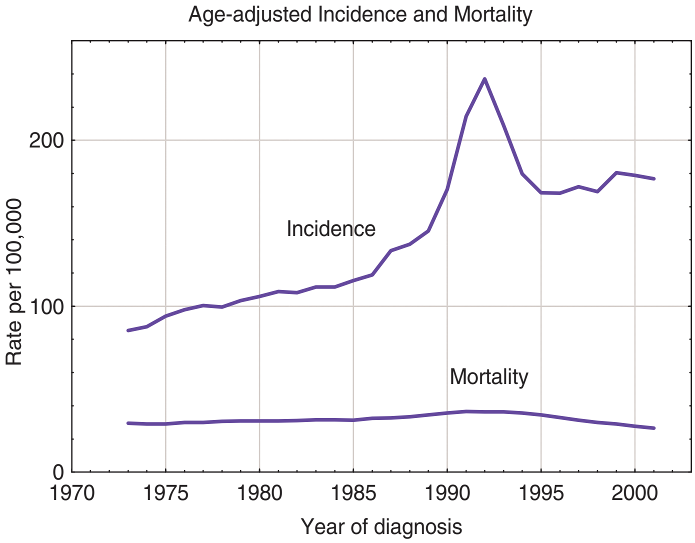


Figure 1. Prostate cancer incidence and mortality rates by year of diagnosis age adjusted to U.S. population in year 2000.

among the key characteristics of the impact of screening on the natural history of the disease that can be identified through latent variable modelling of prostate cancer incidence.

Lead time measures an advance in the diagnosis of prostate cancer due to screening. It adds to the observed survival time even if early detection and treatment were of no benefit.

A large proportion of prostate cancers identified through screening would never be detected in the absence of screening. This phenomenon is called overdiagnosis. Screening brings such cancers to the surface predominantly in the localized stage of the disease leading to an apparent 'favourable' stage-shift. Overdiagnosis has multiple consequences. It leads to over-treatment of men who would never be detected without screening. Also, it modifies apparent estimates of post-treatment survival as over-diagnosed cases appear to be 'cured'. Injection of over-diagnosed cases into the pool of all prostate cancer presentations at diagnosis changes the distribution and the meaning of clinical covariates in men diagnosed with prostate cancer in the PSA era. With the introduction of screening, the prognostic value of such covariates at diagnosis is modified. For these reasons the prognosis for cases diagnosed in the screening era is markedly different than for cases detected naturally. Inclusion of clinical covariates into the model may be misleading and requires special care to adjust the covariate effects for screening patterns in the population.

Modelling represents an important tool for studying screening phenomena. Mathematical and simulation models have been used for inference in cancer screening trials to evaluate controlled randomized screening interventions [6, 7]. While perhaps providing the best design for evaluating the impact of screening, such randomized trials often fail to respond to important challenges. Randomized screening trials typically require decades of observation in order to register a significant effect of screening on cancer mortality. During this period screening modality is often outmoded by new diagnostic advances. Considerable changes in the practice of management of the disease and therapy, often concurrent with advances in screening techniques, make it difficult for screening trials to catch up with an ever evolving scientific and technological progress of cancer detection and treatment.

A number of simulation and analytic models have been designed to translate the results of screening trials into the population setting [8–10, 7]. Given the complexity of the problem at hand no single approach can provide the final answer. In particular, relatively small populations used in screening trials run the risk of not being representative for the national population. Screening patterns operating in the national population are quite different from the artificial homogenized schedules pursued in screening trials. In order to capture the complexity of screening in a population, we adopt an approach of estimation and prediction of the effect of cancer screening directly from population data. We consider screening schedules as a random point process in the population and use characteristics of PSA dissemination to inform the model about the properties of this process. Point estimates and confidence intervals for the model parameters are based on the maximum likelihood technique. Population databases and cancer registries such as SEER provide a unique resource for studying the '*in vivo*' dynamics of the population impact of screening.

2. CANCER INCIDENCE MODEL

2.1. The basic model

We use the classical three-stage model of the natural history of a chronic disease [11]. Prostate cancer is a result of an irreversible transition of the disease through three consecutive stages: disease-free stage, pre-clinical stage and clinical stage. The time spent in disease-free stage is characterized by the age Y (a random variable) at onset of the disease. In the pre-clinical stage disease is asymptomatic and can be detected by a screening test. The duration of the preclinical stage in the absence of screening (a random variable) is termed the sojourn time. If undetected by screening, the disease can either reach the clinical stage or, alternatively, the event of clinical diagnosis gets right censored by a competing risk other than the disease of interest.

It is clear that cancer incidence in an individual is a convolution of two generally dependent survival times. In what follows we will use the notation λ for a hazard function (incidence rate), f for a probability density function (p.d.f.), and G for a survival function (s.f.) unless noted otherwise. Let λ_{O} be the hazard function of $Y =$(age at disease onset). Denote by f_{O} and G_{O} the corresponding p.d.f. and s.f. of Y, respectively.

Prostate cancer incidence $\lambda_{\mathrm{I}}(a,t)$ in year t at the age of a can be written as $\lambda_{\mathrm{I}}(a|t-a)$, where $\lambda_{\mathrm{I}}(a|x)$ is the hazard function for cancer diagnosis at the age of a for a person born in year x. Clearly, for the x-birth cohort

$$\lambda_{\mathrm{I}}(a|x) = \frac{f_{\mathrm{I}}(a|x)}{G_{\mathrm{I}}(a|x)} \tag{1}$$

The functions f_{I} and G_{I} are in fact represented by a fairly complex mixture model which we now start detailing.

Let us condition on the age at tumour onset to obtain the following convolution:

$$f_{\mathrm{I}}(a|x) = \int_0^a f_{\mathrm{I}}(a-y|x,y) f_{\mathrm{O}}(y|x)\,\mathrm{d}y \tag{2}$$

where $f_{\mathrm{I}}(\xi|x,y)$ is a conditional p.d.f. of the random time T from tumour onset to its potential diagnosis. The random variable T represents the duration of the latent disease stage. Generally, $f_{\mathrm{I}}(\xi|x,y)$ is an average over random patterns of screening operating in the population. It is clear that T is a result of two dependent competing risks: the one associated with natural clinical diagnosis through symptoms and the one associated with detection through screening. Dependency between the two risks is a consequence of natural detection and screen-based detection risks sharing the same disease development process in the subject. For example, the event of natural detection indicates a non-zero risk of screen-based detection in the subject as it informs us that the onset of the disease has already happened. This dependency will be modelled through the concept of shared mixed effect (frailty) [12].

In our first approach we identify the onset time Y with the shared mixing variable and make the assumption of conditional independence of potential risks of natural and screen-based detection, given Y. Specifically,

$$G_{\mathrm{I}}(\xi|x,y) = G_{\mathrm{CDx}}(\xi|x,y) G_{\mathrm{SDx}}(\xi|x,y) \tag{3}$$

where G_{CDx} is the s.f. of time to clinical diagnosis (CDx), the sojourn time in the absence of screening, and G_{SDx} is the s.f. of the potential time to screen-based diagnosis (SDx). Here ξ is time since onset, x is date of birth, and y is the age at onset. Note, that G_{SDx} in our model corresponds to a continuous distribution as it is represented as a continuous mixture over random screening schedules in the population.

The density $f_{\mathrm{I}}(\xi|x,y)$ corresponding to cancer diagnosis given birth year x and onset time y can be split into the two crude densities corresponding to the two modes of diagnosis: screening and clinical

$$f_{\mathrm{I}}(\xi|x,y) = f_{\mathrm{SDx}}(\xi|x,y) G_{\mathrm{CDx}}(\xi|x,y) + f_{\mathrm{CDx}}(\xi|x,y) G_{\mathrm{SDx}}(\xi|x,y) \tag{4}$$

With age at tumour onset y integrated out of (4), we obtain a similar relationship describing the partition of the observed p.d.f. $f_{\mathrm{I}}(a|x)$ into the two crude components corresponding to the two modes of detection

$$f_{\mathrm{I}}(a|x) = f^{\mathrm{c}}_{\mathrm{SDx}}(a|x) + f^{\mathrm{c}}_{\mathrm{CDx}}(a|x) \tag{5}$$

where

$$f^{\mathrm{c}}_{\mathrm{SDx}}(a|x) = \int_0^a f_{\mathrm{SDx}}(a-y|x,y) G_{\mathrm{CDx}}(a-y|x,y) f_{\mathrm{O}}(y|x)\,\mathrm{d}y \tag{6}$$

and

$$f^{\mathrm{c}}_{\mathrm{CDx}}(a|x) = \int_0^a f_{\mathrm{CDx}}(a-y|x,y) G_{\mathrm{SDx}}(a-y|x,y) f_{\mathrm{O}}(y|x)\,\mathrm{d}y \tag{7}$$

Note that the crude p.d.f. $f^{\mathrm{c}}_{\mathrm{SDx}}(a|x)$ is a function of age of the *subject* while the net p.d.f. $f_{\mathrm{SDx}}(\xi|x,y)$ conditioned on age at tumour onset is a function of the age of *tumour* $\xi = a - y$. Also, f^{c} here are net densities with respect to death due to causes other than prostate cancer.

It should be noted that the assumption of conditional risk independence is not essential. The logic of this paper can be carried forward with minor changes if G_{SDx} is conditioned on the sojourn time or on a more complex surrogate of natural history process. However, at this point we stop short of making the model more complex and make this and other refinements contingent upon compelling evidence from the data. One class of models consistent with the independence assumption is the one where tumour growth is assumed to be deterministic, and where competing risks of detection are independent, given the tumour growth curve [13].

The distribution for a non-negative random variable can be represented by a survival function G, a hazard function λ, or the p.d.f. f. Dependent on the situation, we will use the most convenient representation and keep in mind that other characteristics can be obtained using the well-known relationships

$$G(t) = \exp\left\{-\int_0^t \lambda(\xi)\,\mathrm{d}\xi\right\} = 1 - \int_0^t f(\xi)\,\mathrm{d}\xi$$

$$f(t) = \frac{\lambda(t)}{G(t)} = -\frac{\mathrm{d}G(t)}{\mathrm{d}t}$$

or their discrete counterparts.

2.2. Modelling cancer detection through screening

This section is devoted to modelling the distribution of potential time to screen-based detection $G_{\mathrm{SDx}}(\xi|x,y)$ conditional on the year of birth x and age at tumour onset y.

For an arbitrary individual from the target population, consider the 'risk' of getting the first screen in his life. Age at first screen may be regarded as a survival time with the instantaneous risk represented by the hazard function $\lambda_{1\mathrm{S}}(a,t)$. Naturally, $\lambda_{1\mathrm{S}}$ depends on age a of the person and the current calendar year t. Generally, it is expected that $\lambda_{1\mathrm{S}}(a,t)$ increases in t starting with the year of PSA introduction. As a function of a, it is reasonable to expect that $\lambda_{1\mathrm{S}}(a,t)$ is increasing initially while the residual life expectancy is still substantial and then decreasing for very old people. An empirical histogram estimate for $\lambda_{1\mathrm{S}}(a,t)$ can be obtained by dividing the number of subjects at the age of a receiving their first screen in year t by the total number of subjects with no evidence of the disease in the (a,t) cell. More precisely, we should count tests in the interval $(t,t+\mathrm{d}t)$ and divide by $\mathrm{d}t$, which results in the same estimate for the grouping interval $\mathrm{d}t = 1$ year. Note that this estimate is inconsistent unless the data are grouped [14].

The evolution of an x-birth cohort up to the age of a can be represented as a line connecting points $(\tau,x+\tau)$, where $\tau\in[0,a]$, on the age by year plane called the Lexis diagram [15]. The probability of no screens by the age of a, $G_{1\mathrm{S}}$, is a survival function obtained by integrating (accumulating) the hazard $\lambda_{1\mathrm{S}}$ over the line

$$G_{1\mathrm{S}}(a|x) = \exp\left\{-\int_0^a \lambda_{1\mathrm{S}}(\tau,x+\tau)\,\mathrm{d}\tau\right\} \tag{8}$$

Denote by $\lambda_{2\mathrm{S}}(a,t)$ the intensity of screening in subjects who already had their first screen. Generally, we expect $\lambda_{2\mathrm{S}}$ to be larger than $\lambda_{1\mathrm{S}}$. Indeed, the fact that the subject has had his first PSA test may identify him as a member of the group that is screened more frequently for reasons such as easier access to secondary testing having done this once already, favourable attitude towards screening in those who choose to have their first test, doctor's recommendations for serial secondary screens following the first one, etc.

The model for risk of diagnosis by cancer screening is based on the following assumptions:

- The probability that a subject born in year x who has never been screened by the age of a receives his first screen in the age interval $(a, a+\mathrm{d}a)$ is $\lambda_{1\mathrm{S}}(a, x+a)\,\mathrm{d}a + o(\mathrm{d}a)$.
- The probability that a subject born in year x who has been screened at least once by the age of a receives a screen in the age interval $(a, a+\mathrm{d}a)$ is $\lambda_{2\mathrm{S}}(a, x+a)\,\mathrm{d}a + o(\mathrm{d}a)$. This assumption defines secondary screens as following a non-homogeneous Poisson process in age with intensity $\lambda_{2\mathrm{S}}(a, x+a)$.
- The probability that a subject born in year x, with the disease onset at the age of y, screened at the age of a is detected with cancer is

$$\begin{array}{ll} 0, & y > a \\ \alpha(a-y), & \text{otherwise} \end{array} \tag{9}$$

  where $\alpha(\xi)$ is the sensitivity of screening, and ξ is the age of tumour at the time of testing. It is natural to specify $\alpha(\xi)$ as an increasing function.

It should be noted that if the whole screening schedule for a person could be considered a realization of a non-homogeneous Poisson process, we would expect $\lambda_{1\mathrm{S}} \equiv \lambda_{2\mathrm{S}}$. This reflects the fact that the time to any next event in such process is characterized by the hazard function equal to the intensity of the process. The fact that $\lambda_{1\mathrm{S}} \not\equiv \lambda_{2\mathrm{S}}$ defies the description of the whole screening schedule for the subject as a non-homogeneous Poisson process. In particular, given the intensity of a non-homogeneous Poisson process, time to the next event depends only on the location of the previous event, but not on the number of events that already happened. In our case it matters whether it is a first or secondary test.

Consider the probability $G_{2\mathrm{SDx}}(\tau|x, a, y)$ that a subject born in year x, with onset of the disease at the age of y who has had his first screen by the age of a is not diagnosed by screening in the age interval $[a, a+\tau]$, $a \geqslant y$. Note that this is a probability of no event in the interval $[0, \tau]$ for a non-homogeneous Poisson process in $\zeta \in [0, \tau]$ with intensity $\lambda_{2\mathrm{S}}(a+\zeta, x+a+\zeta)$ thinned with probability $\bar{\alpha}(\zeta+a-y) = 1-\alpha(\zeta+a-y)$. (We use the notation $\bar{A} = 1-A$ for any A.) The intensity of a Poisson process with intensity λ thinned with probability $\bar{\alpha}$ is given by the product $\lambda\alpha$, so that with $a \geqslant y$,

$$G_{2\mathrm{SDx}}(\tau|x, a, y) = \exp\left\{-\int_0^{\tau} \lambda_{2\mathrm{S}}(a+\zeta, x+a+\zeta)\alpha(\zeta+a-y)\,\mathrm{d}\zeta\right\} \tag{10}$$

If the interval in question is before onset, $a+\tau \leqslant y$, then there is no diagnosis and $G_{2\mathrm{SDx}}(\tau|x, a, y) = 1$. If $a < y$ and $a+\tau > y$, the time interval in ζ where diagnosis is possible starts at $y-a$, so that $G_{2\mathrm{SDx}}(\tau|x, a, y)$ is given by an expression similar to (10) with the lower limit in the integral set at $y-a$. Summarizing, we have

$$G_{2\mathrm{SDx}}(\tau|x, a, y) = \exp\left\{-\int_{\max(y-a,0)}^{\tau} \lambda_{2\mathrm{S}}(a+\zeta, x+a+\zeta)\alpha(\zeta+a-y)\,\mathrm{d}\zeta\right\} \tag{11}$$

where $\int_a^b = 0$ for any $b \leqslant a$.

We are now equipped to derive the probability of no screening diagnosis by the age of $y+\xi$, $G_{\mathrm{SDx}}(\xi|x,y)$ conditional on year of birth x and age at disease onset y, where ξ is time since onset. We have

$$G_{\mathrm{SDx}}(\xi|x,y)=G_{1\mathrm{S}}(y+\xi|x)+\bar{G}_{1\mathrm{S}}(y|x)G_{2\mathrm{SDx}}(\xi|x,y,y)$$
$$+\int_0^{\xi}\bar{\alpha}(v)f_{1\mathrm{S}}(y+v|x)G_{2\mathrm{SDx}}(\xi-v|x,y+v,y)\,\mathrm{d}v \tag{12}$$

The first term in (12) addresses the possibility of no screens by the age of $y+\xi$. The second term addresses the situation when the first screen occurs before onset of the disease at the age of y and no diagnosis is achieved through secondary screens that might happen in the age interval $(y,y+\xi)$. The third term accumulates the probability that cancer is missed at the first and secondary screens occurring after disease onset.

2.3. *Composition of incident cancers*

For an x-birth cohort consider the following random variables that model various potential (other risks removed) durations in the incidence model:

Y, age at onset of the disease;
τ_{CDx}, time from onset to clinical diagnosis;
τ_{SDx}, time from onset to screening diagnosis;
τ_{OC}, age at death due to other causes.

Given x, τ_{OC} is assumed to be independent of Y, τ_{CDx} and τ_{SDx}. Consider the event of cancer diagnosis $\{\mathrm{Dx}|\mathrm{Scr}\}$ in the presence of screening. We may write

$$\{\mathrm{Dx}|\mathrm{Scr}\}=\{Y+\min(\tau_{\mathrm{CDx}},\tau_{\mathrm{SDx}})<\tau_{\mathrm{OC}}\}$$

Cases diagnosed in the presence of screening are composed of two disjoint groups,

$$\{\mathrm{Dx}|\mathrm{Scr}\}=\{\mathrm{CDx}|\mathrm{Scr}\}\cup\{\mathrm{SDx}\}$$

where $\{\mathrm{CDx}|\mathrm{Scr}\}$ is the event of clinical diagnosis through symptoms,

$$\{\mathrm{CDx}|\mathrm{Scr}\}=\{Y+\tau_{\mathrm{CDx}}<\tau_{\mathrm{OC}}\}\cap\{\tau_{\mathrm{SDx}}>\tau_{\mathrm{CDx}}\}$$

and $\{\mathrm{SDx}\}$ is the event of screening diagnosis,

$$\{\mathrm{SDx}\}=\{Y+\tau_{\mathrm{SDx}}<\tau_{\mathrm{OC}}\}\cap\{\tau_{\mathrm{CDx}}>\tau_{\mathrm{SDx}}\}$$

Screen-detected cases are in turn composed of two disjoint groups,

$$\{\mathrm{SDx}\}=\{\mathrm{RDx}\}\cup\{\mathrm{ODx}\}$$

where $\{\mathrm{RDx}\}$ is the event of diagnosis of a *relevant cancer*, and $\{\mathrm{ODx}\}$ is the event of *overdiagnosis*. Relevant cancer is a case diagnosed at a screen such that if screening results on the subject were ignored, he would still be clinically diagnosed later in his life time,

$$\{\mathrm{RDx}\}=\{\tau_{\mathrm{SDx}}<\tau_{\mathrm{CDx}}\}\cap\{\tau_{\mathrm{OC}}>Y+\tau_{\mathrm{CDx}}\} \tag{13}$$

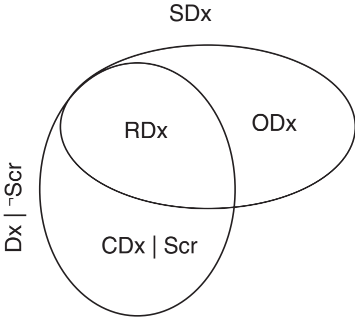


Figure 2. Composition of detected cancers {Dx|Scr}; {SDx}, screen-based diagnosis; {Dx|¬Scr}, cancer diagnosis without screening; {CDx|Scr}, clinical cancer diagnosis in the presence of screening (interval cases); {RDx}, screen-detected cancer that would be detected without screening (relevant case); {ODx}, screen-detected cancer that would not be detected without screening (case of overdiagnosis).

Overdiagnosis is the event of screen detection such that the subject would die from other causes without clinical diagnosis, if the results of screening were ignored

$$\{\text{ODx}\} = \{\tau_{\text{SDx}} < \tau_{\text{CDx}}\} \cap \{Y + \tau_{\text{CDx}} > \tau_{\text{OC}}\} \cap \{Y + \tau_{\text{SDx}} < \tau_{\text{OC}}\}$$

If screening were ignored, detected cancers would be represented by a composition of two disjoint groups, relevant cancers and interval cancers missed by screening,

$$\{\text{Dx}|\neg\text{Scr}\} = \{\text{RDx}\} \cup \{\text{CDx}|\text{Scr}\}$$

The structure of detected cancers described above (see Figure 2) can be verified by elementary sets algebra.

2.4. Overdiagnosis

2.4.1. Overdiagnosis as a long-term outcome. Overdiagnosis as a population measure is defined as a fraction of cancers among all screen detected cancers that would not be detected in the absence of screening. Corresponding to this population statistic is the following conditional probability obtained using the results of Section 2.3:

$$\Pr\{\text{ODx}|\text{SDx}\} = \frac{\Pr\{\text{ODx}\}}{\Pr\{\text{SDx}\}} = \frac{\Pr\{\text{Dx}|\text{Scr}\} - \Pr\{\text{Dx}|\neg\text{Scr}\}}{\Pr\{\text{SDx}\}} \tag{14}$$

It is clear that the above measure represents excess detection due to introduction of screening relative to all screen-detected cases. Alternatively, overdiagnosis and excess detection can be measured relative to all cancer cases.

$$\Pr\{\text{ODx}|\{\text{Dx}|\text{Scr}\}\} = \frac{\Pr\{\text{ODx}\}}{\Pr\{\text{Dx}|\text{Scr}\}} \tag{15}$$

As we will see in Section 5, the two conditional probabilities (14) and (15) show different patterns of cohort effects. While overdiagnosis defined as (14) is decreasing as the cohort life span enters the screening era and screening affects the oldest individuals in the cohort, the conditional probability (15) shows an increasing pattern. The crude probability of cancer diagnosis with screening in place {Dx|Scr} for the x-birth cohort is given by

$$\Pr\{\mathrm{Dx}|\mathrm{Scr}, x\} = \int_0^\infty G_{\mathrm{OC}}(a|x) f_{\mathrm{I}}(a|x)\,\mathrm{d}a \tag{16}$$

where $G_{\mathrm{OC}}(a|x)$ is the survival function modelling death due to other causes for age a and birth year x. Given the hazard function $\lambda_{\mathrm{OC}}(a,t)$ by age and calendar time available for the general population from various sources, G_{OC} can be computed using an expression similar to (8). Likewise, the probability of cancer diagnosis in the absence of screening can be written as

$$\Pr\{\mathrm{Dx}|\neg\mathrm{Scr}, x\} = \int_0^\infty G_{\mathrm{OC}}(a|x) f_{\mathrm{CDx}}(a|x)\,\mathrm{d}a \tag{17}$$

where

$$f_{\mathrm{CDx}} = \int_0^a f_{\mathrm{O}}(y|x) f_{\mathrm{CDx}}(a-y|x,y)\,\mathrm{d}y$$

is the net p.d.f. of age at clinical diagnosis $Y + \tau_{\mathrm{CDx}}$. The denominator of (14) is

$$\Pr\{\mathrm{SDx}|\mathrm{Scr}\} = \int_0^\infty G_{\mathrm{OC}}(a|x) f^{\mathrm{c}}_{\mathrm{SDx}}(a|x)\,\mathrm{d}a$$

where $f^{\mathrm{c}}_{\mathrm{SDx}}$ is the crude density given by (6). Note that overdiagnosis and lead time (see next section) considered as a long-term outcome are a matter of prediction as well as estimation. Their evaluation involves projecting population history beyond the observed box for the duration of a lifetime. For example, a man turning 50 in year 2000 may be diagnosed by screening in year 2004. Whether this man is a case of overdiagnosis depends on the other cause of death risk, trends in disease development and other relevant population dynamics during his potential future lifetime up to the year of, say, 2050. Various other definitions could be proposed to make these characteristics less confounded by future scenarios.

2.4.2. Age-specific presentation. In this subsection we consider a measure of overdiagnosis conditional on age and calendar year at detection. Age-specific conditional p.d.f. of overdiagnosis is proportional to the crude p.d.f. of age at the event of overdiagnosis

$$f(a, \mathrm{ODx}|x) = \int_0^a f_{\mathrm{O}}(y|x) f_{\mathrm{SDx}}(a-y|x,y) \int_0^\infty f_{\mathrm{OC}}(a+s|x) G_{\mathrm{CDx}}(a-y+s|x,y)\,\mathrm{d}y\,\mathrm{d}s \tag{18}$$

the quantity $f(a, \mathrm{ODx}|x)\,\mathrm{d}a$ representing a fraction of the x-cohort overdiagnosed at the age of a to $a + \mathrm{d}a$. Here and in the sequel we omit the subscript to f when it is clear from its arguments what it relates to. Expression (18) represents a formula of total probability averaging over the following sequence of events, in order of the product in (18), onset at the age of y, screening diagnosis at the age of a, and death from other causes s years later prior to clinical diagnosis of prostate cancer. A similar crude p.d.f. for screen-based diagnosis and any diagnosis is given by

$$f(a, \mathrm{SDx}|x) = G_{\mathrm{OC}}(a|x) f^{\mathrm{c}}_{\mathrm{SDx}}(a|x)$$

and

$$f_{\mathrm{Dx}}(a|x) = G_{\mathrm{OC}}(a|x) f_{\mathrm{I}}(a|x)$$

respectively. Finally, age-specific probabilities of overdiagnosis in screen-detected cases and in all prostate cancer cases are given by

$$\Pr\{\mathrm{ODx}|x, a, \mathrm{SDx}\} = \frac{f_{\mathrm{ODx}}(a, \mathrm{ODx}|x)}{f_{\mathrm{SDx}}(a, \mathrm{SDx}|x)} \tag{19}$$

and

$$\Pr\{\mathrm{ODx}|x, a, \mathrm{Dx}\} = \frac{f_{\mathrm{ODx}}(a, \mathrm{ODx}|x)}{f_{\mathrm{Dx}}(a|x)} \tag{20}$$

respectively. Shown in Figure 7 (bottom) are the two age-specific measures of overdiagnosis as estimated from SEER data.

2.5. Lead time

2.5.1. Lead time as long-term outcome. Lead time is the time by which diagnosis of cancer is advanced due to screening in patients who would be detected anyway if screening were not applied. Thus, lead time is defined in the group of patients detected with relevant cancer $\{\mathrm{RDx}\}$ (see (13)). With this definition, we avoid the ambiguity of the lead time in overdiagnosed cases. The group of patients with relevant cancer is characterized by the following history of the disease:

- The subject is born in year x.
- Tumour onset occurs at the age of y. It is not interrupted by death due to other causes.
- Screen-based detection of the tumour occurs at the age of $y + \xi_{\mathrm{SDx}}$. This event is not interrupted by death due to other causes.
- If screening diagnosis were ignored, the tumour would surface at the age of $y + \xi_{\mathrm{CDx}}$ (clinical diagnosis), $\xi_{\mathrm{CDx}} > \xi_{\mathrm{SDx}}$. This event would not be interrupted by death due to other causes.

The lead time is the random variable $\xi_{\mathrm{CDx}} - \xi_{\mathrm{SDx}}$. Its existence is conditional on the membership in the $\{\mathrm{RDx}\}$ group. Formally, the probability of $\{\mathrm{RDx}\}$ for the x-birth cohort is given by

$$\Pr\{\mathrm{RDx}|x\} = \int_0^\infty f_{\mathrm{O}}(y|x) \int_0^\infty f_{\mathrm{CDx}}(\xi|x, y) G_{\mathrm{OC}}(y + \xi|x) \bar{G}_{\mathrm{SDx}}(\xi|x, y)\, \mathrm{d}y\, \mathrm{d}\xi \tag{21}$$

where $\bar{G}_{\mathrm{SDx}}(\xi|x, y) = 1 - G_{\mathrm{SDx}}(\xi|x, y)$ is the net probability of screening diagnosis in ξ years after tumour onset. The product of probabilities in (21) directly follows the definition of relevant cancer (13). We can now write the mean lead time in x-cohort as the conditional expectation

$$E\{\xi_{\mathrm{CDx}} - \xi_{\mathrm{SDx}}|\mathrm{RDx}, x\} = \frac{1}{\Pr\{\mathrm{RDx}|x\}} \int_0^\infty f_{\mathrm{O}}(y|x) \int_0^\infty f_{\mathrm{CDx}}(\xi|x, y) G_{\mathrm{OC}}(y + \xi|x) \times \int_0^\xi f_{\mathrm{SDx}}(\zeta|x, y)(\xi - \zeta)\, \mathrm{d}y\, \mathrm{d}\xi\, \mathrm{d}\zeta \tag{22}$$

Clearly, it is reasonable to limit the time span in the improper integrals by the maximal human lifetime.

2.5.2. Age-specific presentation. An age-specific version of the lead-time can be defined. Consider a conditional mean lead time, given relevant diagnosis at the age of a. The crude joint p.d.f. of age (a) and lead time (s) at relevant diagnosis (RDx) is represented as

$$f_{\mathrm{LT}}(s,a,\mathrm{RDx}|x) = G_{\mathrm{OC}}(a+s|x)\int_0^a f_{\mathrm{O}}(y|x) f_{\mathrm{SDx}}(a-y|x,y) f_{\mathrm{CDx}}(a-y+s)\,\mathrm{d}y \tag{23}$$

where the quantity $f_{\mathrm{LT}}(s,a,\mathrm{RDx}|x)\,\mathrm{d}s\,\mathrm{d}a$ represents a fraction of the x-cohort having relevant diagnosis at the age of a to $a+\mathrm{d}a$ and lead time s to $s+\mathrm{d}s$. The crude age distribution at relevant diagnosis is obtained by integrating the lead time out of (23),

$$f(a,\mathrm{RDx}|x) = \int_0^\infty f_{\mathrm{LT}}(s,a,\mathrm{RDx}|x)\,\mathrm{d}s$$

Now, the mean conditional lead time given age at diagnosis (a) and year of diagnosis ($x+a$) is given by

$$E\{\xi_{\mathrm{CDx}} - \xi_{\mathrm{SDx}}|\mathrm{RDx},x,a\} = \int_0^\infty \frac{f_{\mathrm{LT}}(s,a,\mathrm{RDx}|x)}{f(a,\mathrm{RDx}|x)} s\,\mathrm{d}s \tag{24}$$

2.5.3. Potential lead time. Lead time defined in the previous sections is conditional on the event of screen-detection and on the fact that this is a relevant diagnosis (the potential point of clinical diagnosis at the end of sojourn time occurs before other causes interrupt the natural history). In order to make it a good surrogate of screening dissemination, we now broaden the definition to all cancer cases and consider it in the absence of other causes. The absence of other causes makes the concept of relevant diagnosis obsolete. The lead time for an interval (non-screening) diagnosis is defined as zero. Modifying the argument of the previous section, we have the following expression for the age-specific potential lead time:

$$E\{\xi_{\mathrm{CDx}} - \xi_{\mathrm{SDx}}|x,a\} = \int_0^\infty \frac{f_{\mathrm{LT}}(s,a|x)}{f_{\mathrm{I}}(a|x)} s\,\mathrm{d}s \tag{25}$$

where

$$f_{\mathrm{LT}}(s,a|x) = \int_0^a f_{\mathrm{O}}(y|x) f_{\mathrm{CDx}}(a-y+s|x,y) \begin{Bmatrix} G_{\mathrm{SDx}}(a-y|x,y), & s=0 \\ f_{\mathrm{SDx}}(a-y|x,y), & s>0 \end{Bmatrix} \mathrm{d}y$$

3. LIKELIHOOD

Observed data for the incidence model is represented by the following quantities available by age a and calendar year t in a certain box:

- Population count $P(a,t)$ of people at risk of cancer development.
- Count of cancer cases detected in year t at the age of a, $C(a,t)$.

The conditional likelihood of the data is built as a product of conditional probabilities of cancer detection given the subject is in the risk set for each a, t combination from the box.

$$L = \prod_{a,t} [1 - \lambda_{\mathrm{I}}(a,t)]^{P(a,t)-C(a,t)} \lambda_{\mathrm{I}}(a,t)^{C(a,t)} \tag{26}$$

In the above expression we omit $\mathrm{d}t = 1$ year from the product $\lambda\,\mathrm{d}t$. Taking the log, using the fact that λ_{I} is small, and dropping terms that do not depend on the model parameters, we obtain

$$\ell = \sum_{a,t} C(a,t) \log \lambda_{\mathrm{I}}(a,t) - P(a,t)\lambda_{\mathrm{I}}(a,t) \tag{27}$$

Note that the same likelihood would result if we assumed that C is Poisson distributed with expectation $P\lambda_{\mathrm{I}}$ and that $C(a,t)$ represent independent random variables for different (a,t) pairs (which is not the case in (27)). Maximum likelihood inference is used to obtain point estimates and confidence intervals for the model parameters entering λ_{I}. Maximization of the likelihood can be regarded as minimizing a certain distance between the empirical incidence C/P and its model-based counterpart λ_{I}.

4. SPECIFYING THE MODEL

Since incidence of prostate cancer before the age of 50 is negligibly small, we will associate the birth year x with the year in which the man turns 50. Age variables will be counted out accordingly from this point.

4.1. Age at tumour onset

We use three parametric distribution families in our analysis: Gamma distribution, Weibull distribution and the so-called Moolgavkar, Venzon, Knudson (MVK) distribution [16–18] for the baseline age at tumour onset. The Weibull baseline hazard function is given by

$$h_{\mathrm{O}}(y) = s_{\mathrm{O}} \left(\frac{\Gamma(1 + 1/s_{\mathrm{O}})}{\mu_{\mathrm{O}}} \right)^{s_{\mathrm{O}}} y^{s_{\mathrm{O}}-1}$$

where y is the age past 50. In the above expression Weibull distribution is parameterized through the mean μ_{O} and the shape parameter s_{O} related to the coefficient of variation

$$\sqrt{\frac{\Gamma(1 + (2/s_{\mathrm{O}}))}{\Gamma^2(1 + (1/s_{\mathrm{O}}))} - 1}$$

With the Gamma distribution, we have

$$f_{\mathrm{O}}(y) = \left[\frac{y s_{\mathrm{O}}}{\mu_{\mathrm{O}}} \right]^{s_{\mathrm{O}}} \frac{\mathrm{e}^{-y s_{\mathrm{O}}/\mu_{\mathrm{O}}}}{y\Gamma(s_{\mathrm{O}})}$$

Gamma and Weibull distributions are the two convenience choices.

The MVK distribution [16–18]

$$h_O(y) = \rho AB \frac{e^{(A+B)y} - 1}{B + Ae^{(A+B)y}}$$

where A, B and ρ are identifiable parameters represents a two-stage mechanistic model of carcinogenesis. The MVK distribution has an extra degree of freedom compared to the other two distributions. Weibull distribution showed the best Akaike information criterion (AIC) in the sensitivity analysis.

Included in the model is a trend function $T_O(t)$ that depends on calendar time. This function exerts a multiplicative effect on the baseline hazard so that the hazard of tumour onset depends on age and birth cohort

$$\lambda_O(y|x) = h_O(y) T_O(x + y)$$

The trend is used to model possible changes in the pattern of the disease onset with calendar time due to unspecified factors such as changes in diet, environment and biology of the disease. Note that it is hardly possible to give a biological definition for the tumour onset. From the modelling prospective, tumour onset represents the earliest point in time where cancer could be detected by screening. For this reason changes in detection technology, practice of biopsies for the disease following a positive screens and other diagnostics management issues may also affect the definition. Changes in such practices that are not modelled in a mechanistic fashion are thought of as part of the trend function. We used truncated linear trend functions in data analysis.

4.2. Sojourn time distribution

Sojourn time represents the potential (other risks removed) time from tumour onset to its clinical diagnosis. Weibull distribution with mean μ_{CDx} and shape parameter s_{CDx} is used to model the baseline sojourn time hazard. Two effects can be imposed on the baseline sojourn time distribution:

- *Age dependence*. Sojourn time may be modulated by age for various reasons. Tumour growth biology may depend on the age of the person. Also, tumours developing at a younger age may represent a special subtype that can have different progression characteristics. To model age dependency, the mean sojourn time is regressed on the age at tumour onset y as $\mu_{\mathrm{CDx}} \exp(-\beta_{\mathrm{CDx}} y)$, where the parameter β_{CDx} models correlation between the sojourn time and the onset time.
- *Secular trend*. Sojourn time may be modulated by changes in the practice of cancer detection other than the studied modality of screening. Most notably, before PSA was introduced, prostate cancer was often detected as a result of surgery (transurethral resection of the prostate, TURP) for benign prostate disorders [19]. Other changes in prostate cancer awareness in the population and detection practices may have contributed to a trend of increasing incidence observed before PSA was introduced. These trends in calendar time are modelled using a multiplicative trend function $T_{\mathrm{CDx}}(t)$ acting on the baseline sojourn time hazard.

We have the sojourn time hazard in the form

$$\lambda_{\mathrm{CDx}}(\xi|x, y) = h_{\mathrm{CDx}}(\xi|y) T_{\mathrm{CDx}}(x + y + \xi) \tag{28}$$

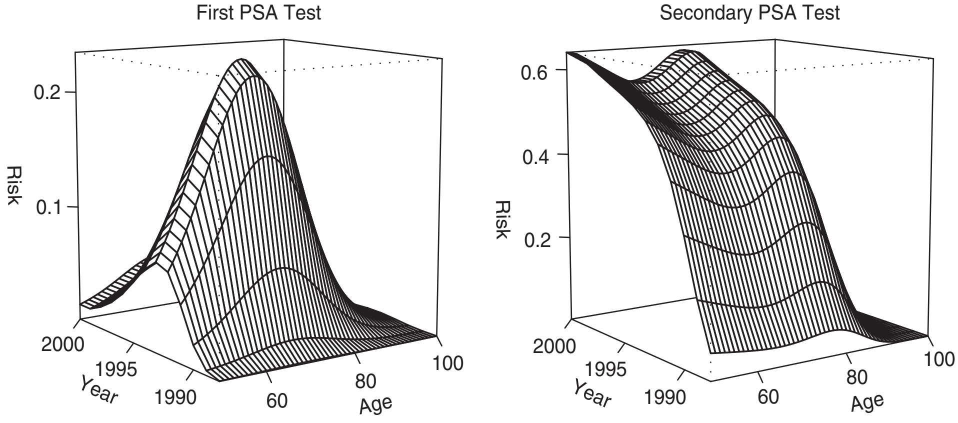


Figure 3. Risks of first λ_{1S} and secondary λ_{2S} PSA tests as estimated from the simulation model by age and calendar year. Left: proportion of never screened men at risk getting their first PSA test. Right: proportion of men screened at least once getting a secondary PSA test.

where x is the birth year, y is age (past 50) at tumour onset, ξ is time since tumour onset, and $h_{\mathrm{CDx}}(\xi|y)$ is Weibull hazard with shape parameter s_{CDx} and mean $\mu_{\mathrm{CDx}}\exp(-\beta_{\mathrm{CDx}}y)$.

4.3. PSA screening model

National Cancer Institute's Statistical Research and Applications Branch has developed a simulator for PSA schedules for arbitrary birth cohorts in the 1916–2000 box. This simulator uses data from the National Health Interview Survey (NHIS) [20] and Surveillance, Epidemiology and End Results (SEER)—Medicare linked database [21]. To extrapolate the data beyond the original age–year box, generalized additive models (R procedure *gam*) were used to smooth the data. A logistic regression model was used for smoothing with the additive main effects of age a and calendar year t represented by thin plate regression splines [22]. No interaction smooth terms were specified. Shown in Figure 3 is an estimate for the risks of first $\lambda_{1S}(a,t)$ and secondary $\lambda_{2S}(a,t)$ PSA tests. It is clear from the figure that the risk of secondary PSA test is several times higher the one for the first test. This observation prompted the development of the two-stage model for screening-based detection described in Section 2.2. Frequency of PSA testing by age increases initially as the man enters the risk zone for prostate cancer. However for the older ages a decreasing pattern is observed perhaps because of limited residual life expectancy and associated diminishing relevance of detection of prostate cancer. Dissemination by calendar year is different for the first and secondary tests. In men who have been screened at least once the frequency increases as PSA is introduced into practice and the surface settles at stable values in the 1990s. The risk of getting the first test by calendar year shows a spike in early 1990s and settles at a lower level later showing a decreasing pattern in the late 1990s. This phenomenon deserves further study. The effect could be a consequence of heterogeneity in people's acceptance of PSA testing. The group of men showing compliance for PSA testing is dissipating with time as such men get tested and leave the set of men 'at risk' for the first test. Another explanation might be that the recent decline in the frequency of new PSA tests is associated with a dissemination of knowledge of various controversial issues surrounding screening for and treatment of prostate cancer.

5. DATA ANALYSIS

SEER database was used to obtain data on more than 350 000 cases of prostate cancer diagnosed in nine areas of the U.S. (San Francisco-Oakland, Connecticut, Detroit, Hawaii, Iowa, New Mexico, Seattle, Utah, Atlanta) as well as population count files corresponding to those cases. We use the modelling box corresponding to age interval [50,85] and calendar year interval [1973–2000]. Age distribution in the U.S. population in year 2000 for men over 50 is used as a standard when age-adjusted characteristics are reported. Risk of death from other causes was derived from the Human Mortality Database [23].

As shown in Figure 1, incidence of prostate cancer before the introduction of PSA showed an increasing trend in calendar time. This is reportedly due to TURP [19], a surgical treatment for a benign enlargement of the prostate. Incidentally, many early stage prostate cancers were discovered as a result of TURPS. With the introduction of PSA modelled from the year of 1987, TURP rates rapidly declined as treatment for benign disease in the prostate was replaced by non-surgical alternatives [19]. In order to model this effect, a linear trend was specified for the sojourn time model (28) for the period 1973–1987, saturating in 1988

$$T_{\mathrm{CDx}(t)} = \begin{cases} 1, & t < 1973 \\ 1 + c(t - 1973), & 1973 \leqslant t \leqslant 1988 \\ 1 + c(1988 - 1973), & t > 1988 \end{cases}$$

The parameter c specifies the slope of the trend. As we are mainly interested in the population effects of PSA screening, the model parameters responsible for pre-PSA trends are considered nuisance parameters. Yet modelling and joint estimation of the pre-PSA era parameters is important as it represents a reference baseline point for the relative effects of PSA. In specifying PSA effects we were looking for a simple trend function that would allow us to provide an adequate description of cancer incidence jointly for the pre- and post-PSA era.

We believed that changes in the onset time distribution over a relatively short observation window are unlikely. Such changes would mean that the prostate cancer has become a different disease over the end of the 20th century. Without compelling evidence we were hesitant to include such changes into the model, particularly since onset time is unobservable. The model with onset trend alone showed a much worse AIC than the model where incidence trend was addressed through sojourn time (AIC 4538 *versus* 510). Although introduction of both trends showed the best AIC of 281, the estimated slope of the onset time trend in this combined model was by two orders of magnitude smaller than that of the sojourn time trend (0.1 *versus* 0.004), which made it too small to be interpreted. Also, the quality of registry data and coding practices have been improving, particularly in the period from 1973 to 1988 where incidence trend was observed. Based on these considerations we decided to proceed without $T_{\mathrm{O}}(t)$ in the model.

The model assigned negative correlation between age at onset and the sojourn time ($\beta_{\mathrm{CDx}} = 0.274$, CI: (0.272,0.279)). This effect is attributable to underestimation of cancer incidence

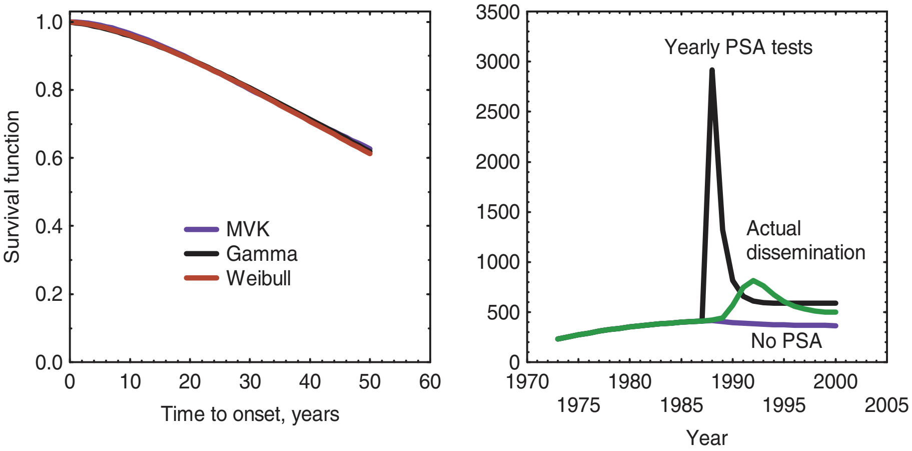


Figure 4. Left: time to tumour onset survival function as estimated using Gamma, Weibull and MVK families of parametric distributions. Time count starts at the age of 50. Right: predictions of prostate cancer age-adjusted (U.S. males in year 2000) incidence under various scenarios of PSA dissemination.

for men over 70 in the early 1970s (Figure 5). Negative correlation is allowing for late onset and early detection thus increasing cancer incidence in the elderly in the absence of PSA. We believe this is an artefact induced by assumed flat secular trends before 1973. Since no data before 1973 was available, our ability to address the issue was limited, and we forced $\beta_{\mathrm{CDx}} = 0$ in the final fit. It should be noted that this left the other model parameters and the shape of the incidence surface past 1980 practically unaltered.

PSA sensitivity as a function of the age of tumour ξ (time since tumour onset) was specified as the following increasing function:

$$\alpha(\xi) = 1 - \exp(-b\xi), \quad b > 0$$

However, when fitting the model, the estimate settled at an 100 per cent sensitive PSA test. The profile likelihood of b is an increasing function (not shown). Therefore, a model with 100 per cent PSA sensitivity was used. With respect to the parameter b, maximum likelihood estimate occurred at the border of the parametric space. It is well known that in such cases likelihood ratio statistic does not generally follow the chi-squared distribution. Basing model selection on the AIC criterion, we used the general rule of thumb that AIC needs to change by more than 2 in order that models be considered as different. This criterion puts the uncertainty in the b parameter at the AIC-confidence interval of $[4, \infty)$. This means that the model indicates that PSA sensitivity reaches half of its maximal value in 2 months after onset or sooner.

It should be stressed that the notion of onset time is a mathematical construct that is difficult to identify with specific *in vivo* biological processes leading to cancer, particularly since it cannot be observed. It might be the case that population data provide limited information to make reliable inference about this unobservable process. With these considerations in mind we conducted sensitivity analyses with respect to the onset time distribution. It turned out

Table I. Estimates of model parameters and confidence intervals.

Parameter	Legend	Point estimate	95% CI
μ_{CDx}	Mean baseline sojourn time	18.558	$(18.345, 18.775)$
s_{CDx}	Shape sojourn time	1.541	$(1.5191, 1.5644)$
c	Slope of trend for sojourn time	0.09354	$(0.09068, 0.09641)$
μ_{O}	Mean age past 50 at tumour onset	72.732	$(72.498, 72.965)$
s_{O}	Shape of age past 50 at tumour onset	1.6153	$(1.6067, 1.6239)$

Time and age is measured in years.

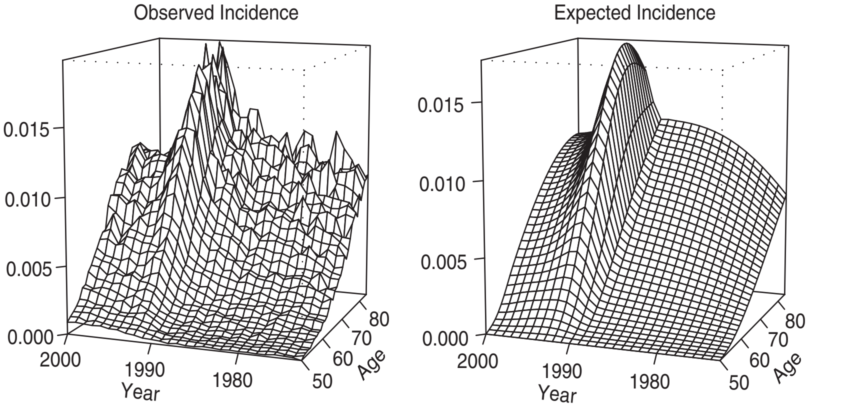


Figure 5. Prostate cancer incidence. Observed (left): empirical estimate of prostate cancer incidence $C(a,t)/P(a,t)$. Expected (right): model-predicted prostate cancer incidence $\lambda_{\mathrm{I}}(a,t)$ by age and calendar year.

that Weibull distribution showed the best AIC (510 for Weibull *versus* more than 1000 for Gamma and MVK). However, we found that all distribution choices provided a very similar estimate for the age at onset survival function (Figure 4). As evident from the figure, the shape of the onset time distribution is fairly unspectacular, and not much flexibility is needed to reproduce the pattern.

Likelihood was maximized by the Powell's method [24] of conjugate directions. Confidence intervals for the model parameters are based on likelihood ratio and inverting of the profile likelihood surface for each parameter. Estimates of model parameters and the corresponding confidence intervals are shown in Table I.

Both sojourn time and onset time are potential times defined as if all other risks were removed. Note that the age at tumour onset goes well beyond the normal human lifetime. This is a consequence of the fact that only a proportion of men would ever develop prostate cancer in their life span. Shown in Figure 5 is a histogram empirical estimate of prostate cancer incidence $C(a,t)/P(a,t)$ and its model-predicted counterpart $\lambda_{\mathrm{I}}(a,t)$ by age and calendar year. The model captures the basic pattern of prostate cancer incidence. The spike effect in the incidence occurring with the introduction of PSA gets more pronounced with age except for very old people. This is a consequence of latent prevalence of the disease accumulating with age.

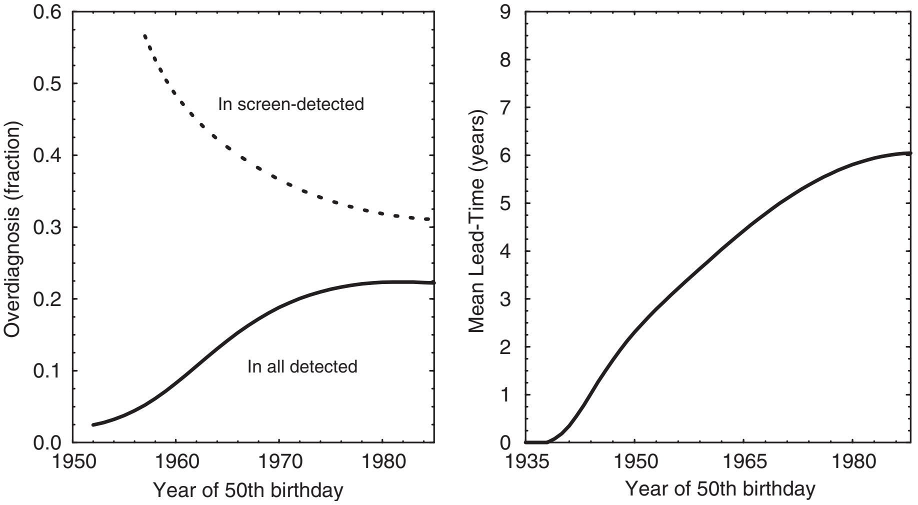


Figure 6. Overdiagnosis (left) and lead-time (right) by birth cohort. Dashed line is the fraction of overdiagnosis in screen-detected patients. Solid line (left) is the fraction of overdiagnosis in all cancer patients.

Shown in Figure 6 is an estimate of lead time and overdiagnosis by birth cohort. Both notions formalized in Sections 2.4 and 2.5 relate to the potential natural history of the disease and population screening exposure over the life span of an individual. As we move the year of birth to the right, more and more of the cohort life span falls on the PSA dissemination era. This leads to an increasing pattern of lead time and overdiagnosis among all detected cancer patients (solid curves). For men entering the age risk zone for prostate cancer at the present time, the model predicts about 6-year mean lead time and 25 per cent overdiagnosis among all detected patients. Interestingly, overdiagnosis in screen-detected cases is a decreasing function of the birth year and settles at about 30 per cent for the present era. Initially for a person born in the 1950s only older ages are affected by PSA dissemination. If detected at such an age, the case is very likely to be overdiagnosed. Indeed, if screening were ignored the disease would have little chance to surface because of the very small expected residual lifetime in older people. This is why the dashed curve in Figure 6 (left) starts high. As we move the potential life history more and more under the PSA exposure, the pool of screen-detected cases gets enriched with relevant cancers that have advanced diagnosis due to PSA yet would surface clinically in their potential residual lifetime if PSA were not applied. Also, since (14) and (15) have the same numerator and in the denominator screen-detected cases represent a subset of all cancer cases, overdiagnosis relative to screen-detected cases (the dashed curve) is always higher than the one relative to all cancer cases (the solid curve). Shown in Figure 7 are the age-specific estimates. These estimates are conditional on cancer diagnosis at the specified age. In screen-detected cases (left part of the figure), lead time and overdiagnosis show a fairly stationary age distribution by calendar year. Bivariate distributions in all detected cases (right part of the figure) follow PSA dissemination pattern showing increasing lead time and overdiagnosis with the introduction of PSA.

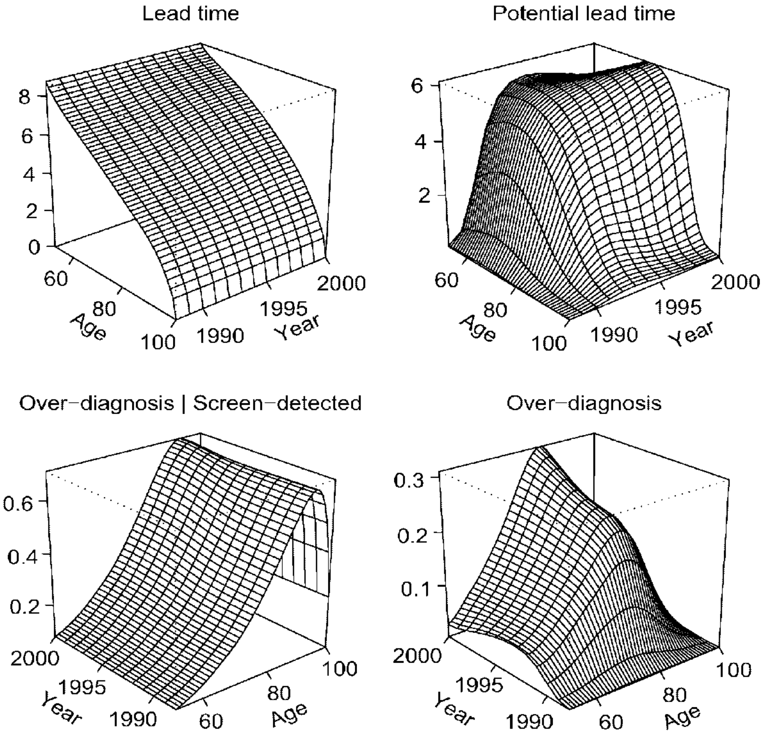


Figure 7. Age-specific fractions of overdiagnosis (bottom) and lead-time (top). Top left shows lead time defined using the competing risk of death due to other causes (22), while top right figure gives mean potential lead time in the absence of other causes (25). Bottom left shows the fraction of overdiagnosis in screen-detected cancer patients (19), and the bottom right figure refers to all cancer patients (20).

Finally, we evaluate two hypothetical scenarios of PSA testing: no PSA screening *versus* yearly PSA tests for every men over 50 starting in 1988. Shown in Figure 4, right, are age-adjusted predictions for the two scenarios as well as the prediction for actual PSA dissemination as estimated by the simulator. This figure gives a range for possible incidence dynamics dependent on PSA dissemination patterns.

6. DISCUSSION

In this paper we have presented a population model of prostate cancer natural history and screening. The model was used to capture the spike of prostate cancer incidence registered after introduction and dissemination of PSA screening for the disease. The link between screening dissemination in the population, natural history of the disease and cancer registration statistics allowed us to predict lead time and overdiagnosis of prostate cancer. The latter two characteristics were rigorously defined in the population setting where screening schedules are random and unknown.

An external data source, NHIS, was used to inform the model about recent frequencies of PSA testing. In a large scale modelling effort, relying on a variety of data sources is unavoidable. However, this may create problems. In particular, there might be differences in PSA dissemination between SEER population (approx 11 per cent of total U.S.) and the NHIS national base. Available data on PSA testing do not allow us to discriminate between a diagnostic and a screening PSA test. Diagnostic tests are usually prompted by symptoms of an enlarged prostate. Prostate enlargement can be caused by locally advanced prostate cancer, or, more likely by a benign disease. Thus cancer discovered as a result of PSA may be a truly screening diagnosis, clinical diagnosis or incidental diagnosis. To prevent misattribution of PSA dissemination, tests performed within 3 months of diagnosis were considered as diagnostic. Use of external data sources introduces additional variability of the estimates that is difficult to control. Bayesian methods might be preferable if this variability is substantial.

Available data do not provide explicit information on unobservable processes such as tumour onset and sensitivity of PSA test prior to diagnosis. Tumour onset can never be observed, and SEER data do not have information on how many men were tested in each particular (age,year) cell. This makes the model estimates of the sensitivity curve and the onset time distribution difficult to verify by the data. Sensitivity is likely to be technology dependent which might introduce a trend in the sensitivity parameters in calendar time.

A number of model applications remained beyond the scope of the present paper. The model can be used to adjust estimates of survival after treatment for the lead-time and overdiagnosis. Regression analysis of prostate cancer survival is confounded by the lead time and overdiagnosis. People diagnosed under less intensive screening generally have shorter lead times and shorter apparent survival times. Also, they are less likely to be over-diagnosed and appear ‘cured’ when followed up for post-treatment failure. In order to adjust estimated treatment effects for such confounding conditional history of the disease given presentation at diagnosis can be considered as frailty when analysing survival data. The population incidence model presented above can be used to derive the frailty distribution as it changes with year of diagnosis, age and clinical covariates. Such model development may help reduce the biases inherent in evaluation of treatment effects using non-randomized tumour registry data.

Jointly, cancer incidence and survival models can be used to build a model of mortality that has a link to PSA dissemination parameters and the effects of treatment and screening. This approach appears promising as a tool to address the reasons for recent declines in prostate cancer mortality.

Analysis of prostate cancer incidence by race indicates that race might be a important variable modulating the natural history of the disease. Reliable estimates are needed for PSA dissemination by race in order to address a possible causal effect of race in prostate cancer.

ACKNOWLEDGEMENTS

This research is supported by National Cancer Institute Cancer Intervention and Surveillance Modelling Network (CISNET) grant U01 CA97414 and Department of Defense grant DAMD 7-03-1-0034. The authors would like to thank R. Etzioni, E. Feuer, A. Mariotto and the broad CISNET community for data resources and many stimulating discussions.

REFERENCES

1. Hsing AW, Devesa SS. Trends and patterns of prostate cancer: what do they suggest? *Epidemiologic Reviews* 2001; **23**:3–13.
2. Hankey BF, Feuer EJ, Clegg LX, Hayes RB, Legler JM, Prorok PC, Ries LA, Merrill RM, Kaplan RS. Cancer surveillance series: interpreting trends in prostate cancerpart I. Evidence of the effects of screening in recent prostate cancer incidence, mortality, and survival rates. *Journal of the National Cancer Institute* 1999; **91**(12):1017–1024.
3. Etzioni R, Legler JM, Feuer EJ, Merrill RM, Chronin KA, Hankey BF. Cancer surveillance series: interpreting trends in prostate cancer—Part III: quantifying the link between population prostate-specific antigen testing and recent declines in prostate cancer mortality. *Journal of the National Cancer Institute* 1999; **91**:1033–1039.
4. Potosky AL, Feuer EJ, Levin DL. Impact of screening on incidence and mortality of prostate cancer in the United States. *Epidemiological Review* 2001; **23**(1):181–186.
5. Surveillance Epidemiology and End Results (SEER) database. http://seer.cancer.gov/.
6. Hu P, Zelen M. Planning clinical trials to evaluate early detection programmes. *Biometrika* 1997; **84**:817–829.
7. Draisma G, Boer R, Otto SJ, van der Cruijsen IW, Damhuis RAM, Schröder FH, de Koning HJ. Lead times and overdetection due to prostate-specific antigen screening: estimates from the European randomized study of screening for prostate cancer. *Journal of the National Cancer Institute* 2003; **95**(12):868–878.
8. Etzioni R, Penson DF, Legler JM, Di Tommaso D, Boer R, Gann PH, Feuer EJ. Overdiagnosis due to prostate-specific antigen screening: lessons from U.S. prostate cancer incidence trends. *Journal of the National Cancer Institute* 2002; **94**:981–990.
9. Zelen M, Lee SJ. Models and the early detection of disease: methodological considerations. *Cancer Treatment and Research* 2002; **113**:1–18.
10. Lee SJ, Zelen M. Statistical models for screening: planning public health programs. *Cancer Treatment and Research* 2002; **113**:19–36.
11. Zelen M, Feinleib M. On the theory of screening for chronic diseases. *Biometrika* 1969; **56**:601–614.
12. Hougaard P. Frailty models for survival data. *Lifetime Data Analysis* 1996; **1**:255–274.
13. Hanin L, Tsodikov A, Yakovlev A. Optimal schedules of cancer surveillance and tumour size at detection. *Mathematical and Computer Modelling* 2001; **33**:1419–1430.
14. Burr D. On inconsistency of Breslow's estimator as an estimator of the hazard rate in the Cox model. *Biometrics* 1994; **50**(4):1142–1145.
15. Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. *Journal of the American Statistical Association* 1983; **78**(381):1–12.
16. Moolgavkar SH, Knudson AG. Mutation and cancer: a model for human carcinogenesis. *Journal of the National Cancer Institute* 1981; 1037–1052.
17. Moolgavkar SH, Luebeck EG. Two-event model for carcinogenesis: biological, mathematical and statistical considerations. *Risk Analysis* 1990; **10**:323–341.
18. Moolgavkar SH, Venzon DJ. Two event model for carcinogenesis: incidence curves for childhood and adult tumours. *Mathematical Biosciences* 1979; **47**:55–77.
19. Merrill RM, Feuer EJ, Warren JL, Schussler N, Stephenson RA. Role of transurethral resection of the prostate in population-based prostate cancer incidence rates. *American Journal of Epidemiology* 1999; **150**(8):848–860.
20. National Health Interview Survey (NHIS). http://www.cdc.gov/nchs/nhis.htm.
21. Surveillance Epidemiology and End Results (SEER)—Medicare linked database. http://healthservices.cancer.gov/seermedicare/.
22. Wood S. Thin plate regression splines. *Journal of the Royal Statistical Society*, *Series B* 2003; **65**:95–114.
23. Human Mortality Database (HMD). //http://www.mortality.org/.
24. Himmelblau DM. *Applied Nonlinear Programming*. McGraw-Hill Book Company: Austin, Texas, 1972.